BOSONIT S. L.

A comprehensive guide outlining the legal and IP frameworks, including compliance checklists and best practices for the collection and processing of medical literature for LLM training

DELIVERABLE 7.1

31/09/2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
AIDESL	Artificial Intelligence Data Extraction of Scientific Literature
CC	Creative Commons
CC0	Creative Commons Zero
CC BY	Creative Commons Attribution
CC BY-NC	Creative Commons Attribution-NonCommercial
CC BY-ND	Creative Commons Attribution-NoDerivatives
CC BY-SA	Creative Commons Attribution-ShareAlike
CDSM Directive	Directive on Copyright in the Digital Single Market
CJEU	Court of Justice of the European Union
D2.1	Deliverable 2.1
EDPB	European Data Protection Board
EU	European Union
GDPR	General Data Protection Regulation
GPAI	General-Purpose Artificial Intelligence
IP	Intellectual Property
ITEA	Information Technology for European Advancement
LLM	Large Language Model
MDR	Medical Device Regulation
MDCG	Medical Device Coordination Group
NLP	Natural Language Processing
OA	Open Access
PETs	Privacy-Enhancing Technologies
SLR	Systematic Literature Review
TDM	Text and Data Mining
WP	Work Package

1 Executive summary

The present deliverable provides a comprehensive legal and intellectual property framework governing the collection, processing, and use of medical scientific literature for the training of large language models within the AIDESL project. Developed under Work Package 2 of this ITEA-funded international research initiative, the document serves as an actionable reference instrument enabling all consortium partners to identify applicable legal obligations, assess the compliance status of prospective data sources, and adopt consistent, auditable decision-making criteria throughout the entire data lifecycle.

The use of scientific literature as AI training data engages multiple concurrent layers of legal protection—article-level copyright, platform-level sui generis database rights, contractual restrictions embedded in publisher terms of service and subscription agreements, and data protection obligations under the GDPR where publications contain identifiable personal data. These regimes stack upon one another, meaning that a single act of large-scale content extraction may simultaneously trigger independent obligations under copyright, database-right, contract, and data protection law. At the core of the analysis lies the interaction between the text and data mining exceptions under Articles 3 and 4 of the CDSM Directive: Article 3 offers a mandatory, contractually non-overridable basis for qualifying research organisations, while Article 4 remains subject to rightsholder opt-outs and publisher licensing restrictions that have now been systematically deployed across the academic publishing sector. The EU AI Act reinforces this landscape by requiring general-purpose AI model providers to implement copyright compliance policies respecting opt-out mechanisms, irrespective of where training occurs. In parallel, the EDPB's Opinion 28/2024 establishes that AI models trained on personal data cannot be presumed anonymous and that development and deployment require independent legal bases—creating a structural tension between the GDPR's data minimisation principle and the AI Act's requirement for representative training datasets.

The deliverable addresses these challenges through a lifecycle-based compliance framework comprising source eligibility criteria, licensing and rights matrices, documentation and traceability requirements, a governance structure with clearly assigned roles and escalation procedures, risk assessments with specific mitigation measures, and practical decision rules for recurrent scenarios. It prioritises openly licensed content (CC0, CC BY) while systematically excluding sources with non-commercial restrictions or valid opt-outs from commercial pipelines. The regulatory landscape remains dynamic—the European Parliament's March 2026 Resolution on Copyright and Generative AI, the pending CJEU reference in Like Company v Google Ireland (C-250/25), and the proposed GDPR reform under the Digital Omnibus package may each materially alter current compliance positions—requiring the consortium's governance framework to incorporate horizon-scanning mechanisms and the capacity to adapt training corpora accordingly. The strategic orientation recommended is founded on three principles: maximising the use of openly licensed content, maintaining strict technical segregation between corpus partitions governed by different licensing regimes, and embedding compliance verification as a structural component of the ingestion pipeline so that no content enters the training corpus without a positively confirmed and recorded legal basis.

2 Introduction and scope

The AIDESL project (Artificial Intelligence Data Extraction of Scientific Literature) is an international research and innovation initiative, co-funded under the ITEA Call 2023 programme, that brings together industry partners, research organisations, and academic institutions from Canada, Germany, Iceland, the Netherlands, and the United Kingdom. The project addresses a well-documented and growing challenge in the medical and life-sciences domain: the manual extraction of data from peer-reviewed scientific publications for the purpose of conducting Systematic Literature Reviews (SLRs) is a labour-intensive, time-consuming, and error-prone process that is becoming unsustainable in the face of the exponential growth of scientific output. AIDESL seeks to overcome this bottleneck by applying large language models (LLMs) and other artificial intelligence techniques to fully automate the extraction of text, tables, figures, and other structured data from published medical literature, with the ultimate goal of reducing the time required to complete an SLR by more than eighty per cent while improving accuracy and reproducibility.

To achieve these objectives, the project is organised into six work packages that follow a cascading structure, beginning with use-case analysis and requirements definition (WP1) and progressing through data management (WP2), development of AI and LLM techniques (WP3), data visualisation and statistical analysis (WP4), integration, demonstration, and evaluation (WP5), and project management, standardisation, exploitation, and dissemination (WP6). Work Package 2 (Data Management), led by Fraunhofer FIT, addresses the foundational layer of the project's data pipeline. Its mandate encompasses the legal and intellectual property dimensions of data collection and processing, the assessment and organisation of medical literature, the design of data curation and storage protocols, and the development of preprocessing techniques tailored to natural language processing applications in the biomedical domain. WP2 produces four sequential deliverables, each corresponding to a distinct task within the work package, and each designed to feed into the technical activities of WP3, WP4, and WP5.

The present document constitutes Deliverable 2.1 (D2.1) of the AIDESL project and corresponds to Task 2.1 (Legal and IP Aspects with Respect to Data Collection and Processing). Its purpose is to provide the consortium with a comprehensive, actionable guide to the legal and intellectual property frameworks that govern the collection, processing, and use of medical scientific literature for the training of large language models. The deliverable is conceived as a reference instrument that enables all project partners to identify the applicable legal obligations, assess the compliance status of prospective data sources, and adopt consistent decision-making criteria throughout the data lifecycle. It is intended to establish a shared legal and operational baseline from which subsequent technical activities—data assessment, curation, storage, and preprocessing—can proceed with adequate legal certainty and traceability.

2.1 Scope of the Document

The scope of this deliverable is defined along three principal axes: the types of content addressed, the categories of data use analysed, and the legal and operational dimensions covered.

With regard to content, the document covers published medical and scientific literature in the forms most relevant to the project's intended training corpus. This includes peer-reviewed journal articles (both full-text and abstracts), systematic reviews, meta-analyses, clinical trial reports, case studies, conference proceedings, and preprints deposited in recognised scholarly repositories. It also addresses non-textual elements embedded within these publications, such as tables, figures, charts, and structured data appendices, insofar as they raise distinct intellectual property considerations. The analysis extends to bibliographic metadata and database records associated with the above categories of literature. The document does not cover grey literature, patient-generated content, electronic health records, clinical datasets obtained directly from healthcare institutions, or proprietary internal datasets held by consortium partners, all of which fall outside the scope of Task 2.1.

With regard to data uses, the document analyses the legal implications of text and data mining (TDM) as defined in Articles 3 and 4 of the Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790), the use of scientific literature as training data for large language models and other AI systems, and the downstream use of trained models for automated data extraction in the context of systematic literature reviews. The analysis considers both research and commercial use scenarios, given the mixed composition of the consortium and the project's dual orientation towards scientific advancement and market impact.

With regard to legal and operational dimensions, the deliverable addresses four interconnected regulatory domains: data protection and privacy (principally under the GDPR and related guidance from the EDPB), copyright and related rights (including the sui generis database right and the TDM exceptions under the CDSM Directive), licensing terms and contractual restrictions imposed by publishers and platform providers, and the regulatory requirements introduced by the EU AI Act (Regulation (EU) 2024/1689) as they pertain to data governance, transparency, and copyright compliance in the context of general-purpose and high-risk AI systems. In addition, the document establishes a governance framework for the consortium, defines roles and responsibilities, provides compliance checklists, and presents practical decision rules for representative data-use scenarios.

Table 1. Scope of D2.1: covered content, covered uses, and exclusions

Dimension	Included in scope	Excluded from scope
Content types	Peer-reviewed journal articles, abstracts, systematic reviews, meta-analyses, clinical trial reports, case studies, conference proceedings, preprints, tables, figures, charts, structured appendices, bibliographic metadata, and database records associated with the relevant literature.	Grey literature, patient-generated content, electronic health records, clinical datasets obtained directly from healthcare institutions, and proprietary internal datasets held by consortium partners.
Data uses covered	Text and data mining under Articles 3 and 4 CDSM, use of scientific literature as training data for LLMs and other AI systems, and downstream use of trained models for automated data extraction in systematic literature reviews.	AI use cases unrelated to the project's data-collection and literature-processing objectives.
Legal and operational dimensions	Data protection and privacy, copyright and related rights, sui generis database right, licensing and contractual restrictions, EU AI Act considerations, governance framework, roles and responsibilities, compliance checklists, and practical decision rules.	Detailed technical implementation guidance beyond the level needed to state legal and governance requirements.
Boundaries with other deliverables and work packages	Legal eligibility criteria, compliance obligations, governance principles, and documentation and traceability expectations relevant to data collection and processing.	D2.2 data assessment methodology; D2.3 storage architecture and operational curation workflows; D2.4 preprocessing pipelines; WP3 model selection and fine-tuning; WP4 visualisation and statistical analysis; WP5 integration and demonstration.

2.2 Boundaries and Exclusions

In order to maintain a clear separation of responsibilities among the deliverables of WP2 and to avoid overlap with subsequent project outputs, the boundaries of this document are defined as follows.

This deliverable does not address the detailed methodologies for the assessment and evaluation of collected data, which constitute the subject matter of Deliverable 2.2 (Data Assessment and Organisation). Accordingly, while the present document establishes the legal criteria that a data source must satisfy in order to be eligible for inclusion in the project's training corpus, the qualitative and quantitative evaluation of the scientific content of that data—including relevance scoring, coverage analysis, and categorisation by medical specialty or research type—is reserved for D2.2.

Similarly, the design and specification of data storage architectures, secure repository infrastructure, dataset documentation standards, and copyright-clearance workflows at the operational level are addressed in Deliverable 2.3 (Data Curation and Storage). The present document confines its treatment of storage and access control to the legal requirements and governance principles that must inform the design of such systems, without prescribing specific technical solutions or infrastructure configurations.

Data preprocessing pipelines, including automated text extraction, data cleaning, NLP-specific transformations, parameterisation strategies, and the development of preprocessing toolkits for medical literature, fall within the scope of Deliverable 2.4 (Data Preprocessing for LLM Training) and are therefore excluded from the present document. Where the legal analysis in this deliverable identifies constraints that bear upon preprocessing activities—for instance, restrictions on the transformation of copyrighted content or requirements for the logging of data transformations—such constraints are stated at the level of legal principle and compliance obligation, without entering into the technical specification of the preprocessing operations themselves.

The selection, benchmarking, and fine-tuning of specific LLM architectures and AI techniques for systematic literature review tasks are the responsibility of WP3 and are not discussed in this deliverable. Likewise, the integration of AI outputs into visualisation and statistical analysis tools (WP4) and the end-to-end demonstration and validation of the project's platform (WP5) lie outside the scope of the present document.

By defining these boundaries explicitly, this deliverable ensures that each component of WP2 addresses a well-delineated set of questions, that downstream deliverables can build on the legal and governance foundations established here without duplication, and that the reader can locate the appropriate reference document for any given aspect of the project's data management strategy.

3 Objectives

The primary objective of this deliverable is to establish a comprehensive legal and intellectual property framework governing the collection, processing, and use of medical scientific literature for the purpose of training large language models within the project. The document is intended to provide all consortium partners with a unified, authoritative reference that identifies the applicable legal and regulatory requirements, translates them into concrete operational guidance, and defines the compliance boundaries within which data-related activities must be conducted. In doing so, the deliverable seeks to ensure that the use of scientific literature throughout the project lifecycle is carried out in full conformity with European Union law, including the General Data Protection Regulation, the Directive on Copyright in the Digital Single Market, the EU AI Act, and all other relevant instruments.

A central aim of the document is to reduce the legal risks associated with the use of copyrighted works, database-protected content, and licensed materials as inputs for AI model development. To this end, the deliverable analyses the scope of copyright and sui generis database rights as they apply to different categories of scientific publications, examines the conditions under which text and data mining exceptions may be lawfully invoked, and delineates the boundary between permitted research uses and activities that require explicit authorisation from rightsholders. By mapping these legal boundaries in advance and translating them into actionable criteria, the document enables the consortium to make informed decisions on source selection and data handling, thereby minimising exposure to infringement claims, contractual breaches, and regulatory sanctions.

The deliverable further aims to provide a structured compliance framework that covers the entire data lifecycle, from initial source identification and acquisition through storage, processing, and preparation for model training, to data sharing, retention, and deletion. This lifecycle-based approach is designed to ensure that compliance is not treated as a one-off assessment but is instead embedded into the operational workflows of the project at every stage. In parallel, the document seeks to establish a common approach to compliance across all consortium partners, harmonising the interpretation and application of legal requirements so that each partner operates under equivalent standards of diligence, regardless of its institutional nature or national jurisdiction.

An equally important objective is to support the governance and accountability requirements that apply to publicly funded research and development projects under CDTI and European Union funding programmes. The deliverable defines roles and responsibilities within the consortium, establishes decision-making and escalation procedures for legally ambiguous situations, and sets out documentation and traceability requirements that enable full auditability of all data-related decisions. These provisions are intended to ensure that the project can demonstrate, at any point during or after its execution, that each dataset incorporated into the training pipeline was selected, acquired, processed, and used in accordance with a documented, legally defensible rationale.

The document also pursues a risk management objective by identifying the principal legal, ethical, regulatory, and reputational risks that arise from the use of medical literature in AI training contexts, and by prescribing specific mitigation measures for each category of risk. This risk-oriented perspective is complemented by the inclusion of practical compliance checklists

and decision rules applicable to recurrent scenarios, such as the use of open-access articles, subscription-based publications, bibliographic metadata, and consortium-shared materials. These tools are designed to enable project teams to resolve routine compliance questions efficiently and consistently, without the need to conduct a full legal analysis for each individual source.

Finally, the deliverable aims to ensure that the technical practices adopted within the project are aligned not only with legal requirements but also with the ethical principles and transparency obligations that underpin the European regulatory framework for artificial intelligence. In particular, the document addresses the data governance, bias detection, and transparency requirements established by the EU AI Act for general-purpose and high-risk AI systems, and provides guidance on how these obligations interact with the copyright and data protection regimes applicable to the training corpus. By integrating legal, regulatory, and ethical dimensions into a single coherent framework, the deliverable provides the consortium with a robust foundation for responsible and compliant AI development in the medical domain.

4 Applicable legal and regulatory framework

The legal assessment of medical scientific literature for LLM training must be understood as a layered exercise, in which multiple legal and regulatory regimes apply concurrently to the same data-use operation. Figure 1 provides a consolidated overview of these interacting layers.

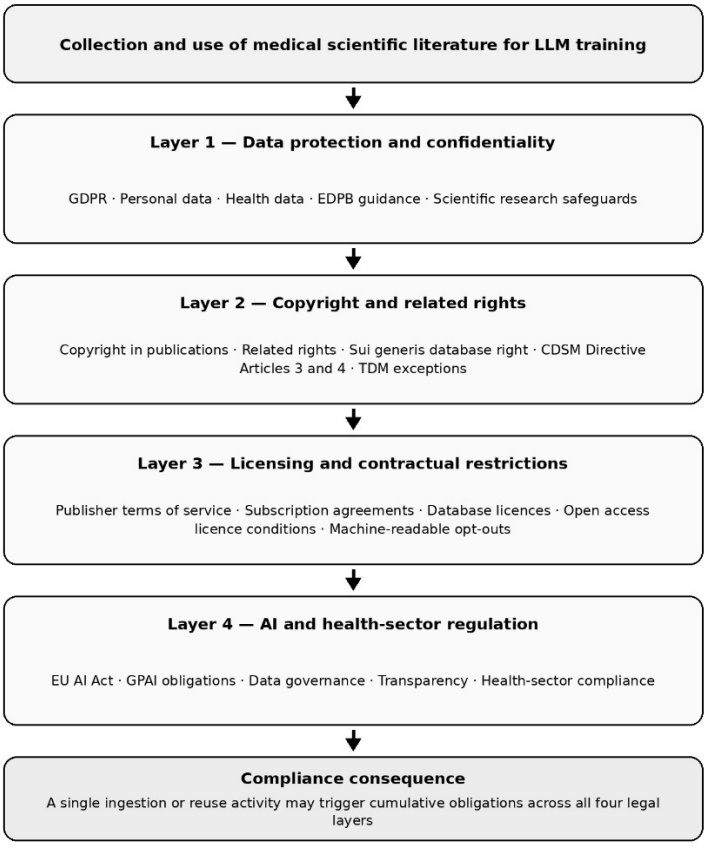

Figure 1. Layered legal framework applicable to the use of medical literature for LLM training

4.1 Data protection and confidentiality

The collection and processing of medical scientific literature for the purpose of training large language models gives rise to significant obligations under the General Data Protection Regulation (Regulation (EU) 2016/679, hereinafter GDPR). Although published scientific articles are, by definition, publicly available, they may nonetheless contain personal data within the meaning of the GDPR. This subsection sets out the principal data protection considerations that apply when such literature is used as training material, including the definition and scope of personal data, the treatment of special categories of data, the applicable legal bases, the relevance of the scientific research exemption, and the technical and organisational measures required to minimise privacy risks.

4.1.1 Personal data in medical literature (Article 4(1) GDPR)

Article 4(1) of the GDPR defines personal data as any information relating to an identified or identifiable natural person. A natural person is considered identifiable when he or she can be identified, directly or indirectly, by reference to an identifier such as a name, an identification number, location data, an online identifier, or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that person.

Medical scientific publications frequently contain personal data, even where the authors have applied de-identification measures. Case reports, clinical trial results, patient demographics, imaging data descriptions, and studies on rare diseases can all embed information that, alone or in combination with other available data, may allow the identification of individual patients. The risks are particularly acute in the case of rare conditions, small patient cohorts, or publications that include detailed demographic or clinical profiles.

For the purposes of this project, the challenge is twofold. First, the medical literature that forms the basis of the training corpus may contain personal data that becomes encoded in model weights during the training process. Second, the resulting AI models may be capable of reproducing or inferring personal information from training data through direct extraction or through responses to targeted queries. Both dimensions must be addressed through appropriate legal bases, technical safeguards, and governance measures.

4.1.2 Special categories of data: health data (Article 9 GDPR)

Health data receives heightened protection under the GDPR as a special category of personal data. Article 9(1) establishes a general prohibition on the processing of data concerning health, among other sensitive categories. This prohibition applies unless one of the specific exceptions enumerated in Article 9(2) is met.

Among the exceptions most relevant to this project, Article 9(2)(j) permits the processing of special categories of data, including health data, when such processing is necessary for scientific research purposes. This exception is subject to the requirement that appropriate safeguards be in place, in accordance with Article 89(1). Such safeguards must ensure respect for the principle of data minimisation and may include pseudonymisation, provided that the research purposes can still be fulfilled in that manner.

In addition to the GDPR, the EU AI Act (Regulation (EU) 2024/1689) introduces a complementary exception in its Article 10(5). This provision allows providers of AI systems to process special

categories of data, including health data, to the extent strictly necessary for the purpose of detecting and correcting biases. However, this exception is subject to six cumulative conditions: pseudonymisation of the data, implementation of appropriate access controls, deletion of the data once the correction has been achieved, documentation of the processing in the relevant records, a demonstrated link between the bias and the special category data, and the absence of less intrusive alternatives.

4.1.3 EDPB Opinion 28/2024: implications for AI model training

The European Data Protection Board (EDPB) adopted Opinion 28/2024 on 17 December 2024, addressing four critical questions concerning the relationship between AI models and personal data protection. This opinion represents a watershed moment in the interpretation of the GDPR as applied to AI development and deployment, and has direct implications for any project that involves training language models on data that may include personal information.

First, the EDPB established that AI models trained on personal data cannot, in all cases, be considered anonymous. The Board rejected the assumption that the aggregation of individual data points into model parameters automatically renders those data anonymous. For a model to qualify as anonymous, both the likelihood of direct extraction of personal data from the model and the likelihood of obtaining personal data through targeted queries must be demonstrated to be insignificant. This determination must be made on a case-by-case basis, taking into account the architecture of the model, the nature of the training data, and the safeguards applied.

Second, the EDPB confirmed that the legitimate interest of the controller, within the meaning of Article 6(1)(f) of the GDPR, can serve as a valid legal basis for AI model training. However, this legal basis is available only where the controller satisfies a three-step cumulative test: the identification of a concrete and specific legitimate interest, the demonstration that the processing is strictly necessary to pursue that interest, and a balancing exercise showing that the interest is not overridden by the fundamental rights and freedoms of the data subjects. The EDPB further noted that this assessment must be documented and must take into account the reasonable expectations of the data subjects.

Third, the opinion underscored that the development and deployment phases of an AI model constitute distinct processing operations, each of which requires an independent legal analysis. A legal basis established for the training phase does not automatically extend to subsequent deployment. Controllers must therefore ensure that a valid legal basis exists for each phase of the model lifecycle.

Fourth, the EDPB addressed the question of whether unlawful processing during the development phase can be remedied retrospectively through anonymisation of the resulting model. The Board acknowledged the theoretical possibility but set a high evidentiary bar, requiring controllers to demonstrate, through rigorous technical assessment, that the model does not retain extractable personal data and cannot be used to derive personal information through inference.

4.1.4 Commission's proposed GDPR reform for AI training

In November 2025, the European Commission proposed a significant amendment to the GDPR as part of the EU Digital Omnibus package. The proposal would codify legitimate interest as a recognised legal basis for AI model training, moving beyond the current situation in which national supervisory authorities have applied divergent interpretations. If adopted, this reform would represent the most substantial modification to the GDPR since its entry into application,

providing harmonised legal certainty for the development of AI systems across the European Union.

It is important to note, however, that even under the proposed reform, the processing of special categories of personal data, including health data, would remain subject to the heightened safeguards of Article 9. The legitimate interest basis alone would not be sufficient to justify the processing of health data for AI training; controllers would continue to need to rely on one of the specific exceptions in Article 9(2), such as the scientific research exemption under Article 9(2)(j).

4.1.5 The scientific research exemption

The GDPR establishes a special legal regime for the processing of personal data for scientific research purposes, which eases certain restrictions that would otherwise apply. This regime is distributed across several provisions of the Regulation and is of particular relevance to projects such as this one, where medical literature containing personal data is processed for the development of AI tools with a research orientation.

Article 5(1)(b) of the GDPR sets out the principle of purpose limitation, under which personal data must be collected for specified, explicit, and legitimate purposes and not further processed in a manner incompatible with those purposes. However, the same provision establishes a presumption of compatibility for further processing carried out for scientific research purposes, in accordance with Article 89(1). This presumption significantly broadens the scope of lawful reuse of data originally collected for other purposes.

Article 9(2)(j) permits the processing of special categories of data, including health data, when such processing is necessary for scientific research purposes. This exception is conditional on the implementation of appropriate safeguards, which, pursuant to Article 89(1), must ensure respect for the principle of data minimisation. Pseudonymisation is identified as a preferred measure, provided that the research objectives can still be achieved.

Furthermore, Articles 15 to 21 of the GDPR confer a range of rights on data subjects, including rights of access, rectification, erasure, restriction of processing, and the right to object. However, Article 89(2) allows Member States to derogate from these rights when personal data are processed for scientific research purposes, to the extent that such rights would render impossible or seriously impair the achievement of the research objectives. The availability and scope of these derogations vary across national transpositions and must therefore be assessed on a jurisdiction-by-jurisdiction basis.

Recital 33 of the GDPR acknowledges that it is often not possible to fully identify the purpose of personal data processing for scientific research at the time of data collection, and therefore allows for broader consent formulations in the research context.

Whether the training of large language models qualifies as scientific research under the GDPR is not yet settled as a matter of law. A research organisation training a model for non-commercial scientific purposes will typically have a stronger claim to the research exemption than a commercial AI developer training a general-purpose model. For this project, the classification will depend on the specific purposes of the model, the institutional context of the consortium partners, and the degree to which the training activities are oriented towards the advancement of scientific knowledge rather than commercial exploitation.

Table 2. GDPR legal bases relevant to AI training on medical literature

Legal mechanism	When it may apply	Main conditions
Article 6(1)(f) GDPR – Legitimate interest	Where personal data contained in medical literature are processed for AI model development and the controller relies on a legitimate interest	A concrete and specific legitimate interest must be identified; the processing must be necessary; and a balancing test must show that the interest is not overridden by the rights and freedoms of the data subjects
Article 9(2)(j) GDPR – Scientific research exemption	Where medical literature includes health data or other special-category data and processing is necessary for scientific research purposes	The processing must be necessary for scientific research; Article 89(1) safeguards must be implemented; data minimisation must be respected; pseudonymisation should be used where possible
Article 89(1) GDPR – Safeguards for scientific research	Where personal data are processed for scientific research purposes	Appropriate technical and organisational safeguards must be implemented, including measures such as pseudonymisation, minimisation, access restrictions, and governance controls
Article 5(1)(b) GDPR – Research compatibility principle	Where previously collected personal data are further processed for scientific research purposes	Further processing for scientific research is presumed compatible if carried out in accordance with Article 89(1) safeguards

Table 3. Limitations and practical relevance of the main GDPR-related mechanisms

Legal mechanism	Main limitations	Relevance for AIDESL
Article 6(1)(f) GDPR – Legitimate interest	It does not by itself authorise the processing of special categories of personal data such as health data; it must be assessed separately for development and deployment phases	Potentially relevant for certain processing operations involving personal data in published literature, but insufficient on its own where health data are involved
Article 9(2)(j) GDPR – Scientific research exemption	Whether LLM training qualifies as scientific research depends on the specific purpose, institutional context, and degree of commercial orientation	Highly relevant where the corpus includes medical publications containing health-related personal data
Article 89(1) GDPR – Safeguards for scientific research	Safeguards do not remove the need for a valid legal basis; they	Relevant as a cross-cutting compliance layer for research-

	must be tailored to the actual risks and the nature of the data	oriented processing involving personal data
Article 5(1)(b) GDPR – Research compatibility principle	The research qualification is not automatic; compatibility does not eliminate the need to assess lawfulness, necessity, and proportionality	Relevant for assessing the secondary use of personal data embedded in published scientific literature
AI Act Article 10(5) – Bias detection and correction exception	It is a narrow exception; it does not authorise broad or general-purpose health-data processing for model training	Relevant only in limited and specific circumstances, particularly if downstream systems fall within high-risk AI use contexts
Anonymisation claim after training	The EDPB rejects the assumption that training automatically anonymises data; the evidentiary threshold is high	Highly relevant because the model may retain or reproduce personal information

4.1.6 Anonymisation and its limits in the context of AI models

A recurring assumption in the field of AI development is that the training process itself anonymises personal data by dispersing it across millions of model parameters, such that individual data points can no longer be extracted. The EDPB's Opinion 28/2024 has decisively rejected this assumption. The Board established that anonymisation is not an automatic consequence of training; rather, it must be affirmatively demonstrated through technical evidence.

For a model to be considered anonymous, two conditions must be met simultaneously: the likelihood of direct extraction of personal data from the model must be insignificant, and the likelihood of obtaining personal data through queries submitted to the model must also be insignificant. The assessment must take into account the state of the art in extraction and inference techniques, which are evolving rapidly. Consequently, a model that may appear anonymous today could become re-identifiable as new attack methods emerge.

The EDPB further acknowledged that unlawful processing during the development phase may, in principle, be remedied through subsequent anonymisation of the model. However, the Board set a high evidentiary bar for this pathway. Controllers must provide robust technical evidence that the model does not retain personal data in any extractable form and cannot be used to derive personal information through targeted interaction. This requirement has significant practical implications for the design of training pipelines, the selection and preprocessing of training corpora, and the evaluation of trained models prior to deployment.

4.1.7 Re-identification risks in medical data

Recent research has highlighted that traditional de-identification techniques applied to medical data may not provide robust protection against re-identification when that data is subsequently processed by AI systems. Studies have demonstrated that AI-enabled attacks can identify individuals from anonymised chest X-rays, electrocardiograms, magnetic resonance images, and

even gait data. These findings indicate that AI systems themselves can reverse the protections that were applied during data preparation.

This creates a specific challenge for the present project. Medical literature intended as training material may include de-identified clinical data that, while compliant with conventional anonymisation standards, remains vulnerable to re-identification by the very models being trained. The consortium must therefore adopt a layered approach to privacy protection, combining legal compliance measures with technical safeguards that account for the specific capabilities of AI systems to re-identify individuals from ostensibly anonymised data.

4.1.8 Privacy-enhancing technologies

Privacy-enhancing technologies (PETs) are increasingly recognised as necessary complements to legal compliance measures when personal data is processed for AI development. These technologies provide practical mechanisms to reduce re-identification risks, limit data exposure, and support compliance with the data minimisation and security requirements of the GDPR and the data governance obligations of the EU AI Act.

Federated learning enables model training across distributed datasets held by different institutions without requiring the centralisation of sensitive data. Under this approach, each participating node trains a local model on its own data and shares only aggregated model updates, rather than raw data, with a central server. This architecture reduces the exposure of personal data and aligns with the GDPR's data minimisation principle, although it does not eliminate all privacy risks, as model updates may still leak information about the underlying training data.

Differential privacy introduces mathematically calibrated noise into the training process to prevent the extraction of information about individual data points. When properly implemented, differential privacy provides a formal guarantee that the inclusion or exclusion of any single record in the training dataset has a negligible effect on the model's outputs. This technique is particularly relevant where the training corpus includes data derived from identifiable individuals.

Synthetic data generation creates artificial datasets that preserve the statistical properties of the original data without containing real personal information. When used as a substitute for, or complement to, real medical literature in the training pipeline, synthetic data can reduce the privacy footprint of the model while maintaining its analytical utility. However, the quality and fidelity of synthetic data must be carefully validated to ensure that it does not introduce bias or reduce model performance.

Homomorphic encryption enables computation on encrypted data without requiring decryption. While this technology remains computationally intensive and is not yet widely deployed in large-scale model training, it represents a promising avenue for enabling privacy-preserving processing of sensitive medical data in future iterations of the project.

The consortium should evaluate the applicability of these technologies at each stage of the data lifecycle, taking into account the nature of the data being processed, the privacy risks identified, and the technical feasibility of implementation within the project's infrastructure and timeline.

4.2 Copyright and related rights

4.2.1 CDSM Directive 2019/790: Text and Data Mining Exceptions

The EU Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790, hereinafter the *CDSM Directive*) constitutes the primary legislative instrument governing the intersection between copyright law and text and data mining (TDM) activities within the European Union. For the purposes of this deliverable, its Articles 3 and 4 are of central importance, as they establish two distinct exceptions that determine the conditions under which copyrighted works may be used for TDM, including in the context of AI model training.

Article 3 of the CDSM Directive provides a mandatory and imperative exception for TDM carried out for the purpose of scientific research. This exception is limited in its personal scope to research organisations and cultural heritage institutions, and it cannot be overridden by contractual provisions. Rightsholders are unable to opt out of this exception, which means that, provided the conditions are met, qualifying entities may reproduce and extract content from lawfully accessed works without prior authorisation. The exception also permits the retention of copies for the purposes of scientific verification. The mandatory nature of Article 3 reflects the European legislator's intent to safeguard non-commercial scientific inquiry from being impeded by copyright restrictions.

Article 4 introduces a broader, general-purpose TDM exception available to any natural or legal person, including commercial entities. Unlike Article 3, however, this exception is subject to a rightsholder opt-out mechanism: under Article 4(3), rightsholders may reserve their TDM rights by expressing such reservation in a machine-readable manner. Where such an opt-out has been validly exercised, the general TDM exception ceases to apply and any reproduction or extraction of the reserved works requires a licence or other authorisation. As with Article 3, lawful access to the works remains a prerequisite.

4.2.2 The TDM Exceptions in Practice

The practical application of Articles 3 and 4 raises several significant issues for AI training workflows. With respect to the opt-out mechanism under Article 4(3), the manner in which rightsholders may validly reserve their TDM rights has become a matter of increasing practical and legal importance. The standard technical mechanism currently relied upon is the robots.txt protocol, through which website operators can signal restrictions on automated access and data extraction. However, the Hamburg Regional Court has discussed whether natural-language restrictions—such as those found in publisher terms of service—may also qualify as "machine-readable" in light of the capacity of modern AI systems to process and interpret unstructured text. This question remains open, and its resolution may have significant implications for the scope and enforceability of opt-outs.

In practice, most major academic and scientific publishers now include explicit TDM and AI training restrictions in their terms of service. Publisher subscription agreements and database licences (e.g., those governing access to Scopus, Web of Science, and the licensed full-text content of PubMed) typically contain clauses that restrict automated downloading, systematic extraction, and data mining beyond what the TDM exceptions would otherwise permit. These contractual restrictions operate independently of the copyright exceptions and may impose additional limitations on the consortium's ability to use such content for LLM training purposes, particularly where such use extends beyond what the relevant licence contemplates.

4.2.3 European Parliament Resolution on Copyright and Generative AI (March 2026)

On 10 March 2026, the European Parliament adopted a Resolution on Copyright and Generative AI which, although non-binding, signals a potentially significant shift in the EU's policy approach to the use of copyrighted works in AI training. The Resolution proposes, among other measures, the introduction of a flat-rate licensing fee calculated as a percentage of global turnover (with suggested values in the range of five to seven percent) to compensate creative industries for the use of their works in AI training. It further suggests extending the territorial reach of EU copyright law to cover AI systems offered in the European market, regardless of where the training takes place, and calls for enhanced transparency requirements beyond those already established under the AI Act, with the European Union Intellectual Property Office (EUIPO) potentially serving as a monitoring body.

While this Resolution does not have the force of law, it reflects a growing legislative consensus that the TDM exceptions in the CDSM Directive were not designed to accommodate the scale and nature of large-scale generative AI training. The consortium should monitor the legislative developments that may follow from this Resolution, as any resulting legislation could retroactively affect the legal basis for the use of copyrighted materials in the project's AI training activities.

4.2.4 Research Versus Commercial Use

The distinction between Articles 3 and 4 of the CDSM Directive has direct and material consequences for the consortium's activities. A university research group or a recognised research organisation that trains a medical AI model for the purpose of scientific research may invoke the mandatory TDM exception under Article 3, which cannot be overridden by contract and against which no opt-out is available to rightsholders. By contrast, a commercial AI company seeking to train a general-purpose language model cannot rely on Article 3 and must instead depend on the general TDM exception under Article 4, which is subject to the opt-out mechanism and to publisher licensing restrictions.

The key differences between the two exceptions may be summarised as follows. Article 3 establishes a mandatory exception that is imperative in nature, meaning it takes precedence over conflicting contractual terms. Its scope is restricted to research organisations and cultural heritage institutions, and it is limited to activities carried out for the purpose of scientific research. No opt-out mechanism is available to rightsholders. Article 4, by contrast, is available to a broader range of actors, including commercial entities, but is qualified by the right of rightsholders to reserve their TDM rights through machine-readable declarations. It is also subject to publisher restrictions and licensing agreements, which may further limit the scope of permissible use.

Whether AI model training qualifies as "scientific research" within the meaning of Article 3 remains a contested and unresolved question. A research organisation that trains a model for non-commercial scientific purposes—such as the development of diagnostic support tools or the analysis of biomedical literature—has a stronger claim to the protection of Article 3 than a commercial developer training a general-purpose large language model. The consortium must carefully assess the nature and purpose of each training activity in order to determine which exception, if any, may apply.

4.2.5 Landmark Court Decisions

Hamburg Regional Court and Higher Regional Court

In a landmark decision, the Hamburg Regional Court (2024) ruled that the creation of a dataset containing copyrighted content for the purpose of AI training constitutes lawful TDM within the meaning of the CDSM Directive. The court rejected a restrictive interpretation that would have excluded AI training datasets from the scope of TDM, reasoning that the European legislator itself had envisaged AI training as falling within TDM, as evidenced by the AI Act's references to the opt-out mechanism under Article 4(3) CDSM. In the case at hand, the court applied the scientific research exception under Article 3 CDSM (transposed into German law as Section 60d of the German Copyright Act) because the defendant was a research organisation that made the resulting dataset freely available for research purposes.

On appeal, the Higher Regional Court of Hamburg (2026) introduced a more nuanced analytical framework by distinguishing three distinct phases of the AI training lifecycle. The first phase, consisting of the copying and compilation of data for dataset creation, was found to be permissible under the TDM exception. The second phase, involving the actual training process by which the model internalises patterns from the copyrighted material, was identified as potentially constituting a separate act of reproduction that may require its own independent legal basis. The third phase, relating to the generation of outputs by the trained model, was characterised as a further act of reproduction where such outputs contain or substantially reproduce protected elements of the original works. Critically, the appellate court held that memorisation during the training process—whereby the model retains and is capable of reproducing substantial portions of protected works—could constitute copyright infringement not covered by the TDM exception.

Munich Regional Court – GEMA v. OpenAI (November 2025)

In the case of GEMA v. OpenAI, decided by the Munich Regional Court in November 2025, the court concluded that certain forms of AI training involving protected musical works constituted copyright infringement and that the TDM exception did not apply in the circumstances. This decision is particularly significant because it represents one of the first instances in which a German court has held that specific AI training practices fall outside the scope of the TDM exceptions, reinforcing the view that the applicability of the exception depends on the specific facts and circumstances of each case, including the nature of the works involved, the purpose of the training, and the manner in which the works are used.

EU Versus US: Copyright Approaches to AI Training

Although US law is not directly applicable to the consortium's activities, a comparative overview of the US approach provides useful context for understanding the distinct nature of the EU copyright regime. In the United States, courts in cases such as *Bartz v. Anthropic* and *Kadrey v. Meta* have held that feeding copyrighted texts into AI models may constitute a transformative use that does not directly compete with the original works, and may therefore qualify as fair use under US copyright law. The US Copyright Office, in its May 2025 report, adopted the position that training may constitute prima facie infringement of reproduction rights, while acknowledging that the application of fair use is fact-specific and a matter of degree.

The fundamental difference between the two legal systems is that the EU does not recognise a general fair use doctrine. The EU operates under a closed system of specific exceptions and limitations to copyright, meaning that any use of protected works not explicitly covered by a statutory exception requires prior authorisation from the rightsholder. For the consortium, this

means that commercial AI training involving copyrighted medical literature must be justified either by a licence or by the TDM exceptions under Articles 3 or 4 of the CDSM Directive. There is no residual open-ended defence comparable to US fair use that can be invoked when neither a licence nor a statutory exception applies. This closed-list approach makes the correct characterisation of the consortium's activities—as scientific research or as commercial use—a decisive factor in determining the applicable legal framework.

Table 4. Comparison of Articles 3 and 4 CDSM for text and data mining and AI training.

Criterion	Article 3 CDSM	Article 4 CDSM
Eligible actors	Research organisations and cultural heritage institutions	Any natural or legal person, including commercial entities
Purpose	Scientific research	General-purpose text and data mining
Mandatory nature	Mandatory and imperative exception	Non-mandatory in practice because it is subject to rightsholder reservation
Contract override	Contractual restrictions cannot override the exception	Contractual and opt-out restrictions remain highly relevant
Rightsholder opt-out	No opt-out available	Opt-out possible under Article 4(3)
Lawful access requirement	Yes	Yes
Commercial use	Generally not the core target of the provision	Yes, potentially applicable to commercial actors
Main legal advantage	Stronger protection for qualifying research entities	Wider personal scope
Main legal limitation	Narrow personal and functional scope	Can be neutralised by valid opt-outs and licensing restrictions
Relevance for AI training	More favourable where model training genuinely qualifies as scientific research	More relevant for commercial or mixed-use AI training scenarios
Main compliance risk	Whether the activity truly qualifies as scientific research	Whether the source is subject to opt-out, licence, or contractual restriction

4.3 Licensing terms and contractual restrictions

The use of medical scientific literature for the purpose of training large language models (LLMs) is not exclusively governed by copyright law and its statutory exceptions. It is equally shaped by a complex layer of licensing agreements, contractual arrangements, and machine-readable opt-out mechanisms that operate alongside, and sometimes in tension with, the rights afforded by the Text and Data Mining (TDM) exceptions under the Directive on Copyright in the Digital Single Market (CDSM Directive, EU 2019/790). This section analyses the principal modalities through which rightsholders exercise control over AI training uses: commercial licensing to AI developers, the structural implications of copyright transfer agreements, opt-out mechanisms, the specific constraints imposed by open-access licence conditions, and the legal interaction between contractual restrictions and mandatory TDM exceptions.

4.3.1 The Commercial Licensing Landscape

Academic publishers have increasingly entered into direct commercial licensing agreements with AI developers, effectively monetising their content repositories as training data. This emerging market was valued at approximately USD 368 million in 2024 and is projected to reach USD 2.9 billion by 2033, reflecting the strategic value that high-quality, domain-specific corpora — including biomedical and clinical literature — hold for the development of specialised AI models.

Several significant deals have been publicly disclosed. Taylor & Francis entered into an agreement with Microsoft valued at USD 10 million for AI training access to its academic content. Wiley recorded USD 23 million in one-time revenue from a licensing agreement with an undisclosed technology company. Springer Nature has similarly entered licensing arrangements, though the financial terms have not been disclosed publicly. Elsevier, by contrast, has adopted a more cautious stance, preferring to observe the evolving market while simultaneously developing proprietary AI-driven research tools that leverage its own content holdings.

These transactions are significant from both a legal and a practical standpoint. They demonstrate that publishers regard AI training use as a commercially licensable right, independent of whether such use would otherwise fall within a statutory TDM exception. For entities seeking to train models on licensed medical literature, the existence of such agreements establishes a clear market for the licences in question, which courts and regulators may take into account when assessing whether unauthorised use causes market harm to the rightsholder.

4.3.2 Copyright Transfer Agreements and the Author Problem

A structural feature of academic publishing that has direct legal consequences for AI licensing is the widespread practice of comprehensive copyright transfer. Standard publishing contracts in the academic sector frequently include broad assignment clauses under which authors transfer all copyright ownership in their works — in all forms and media — to the publisher. As a result, publishers acquire the legal standing to sublicense their accumulated content for AI training purposes without the need to seek the consent of the original authors, and without any obligation to share the resulting licensing revenues with them.

This arrangement has attracted criticism from authors' rights advocates, who argue that many academic authors are unaware that their published works may be incorporated into AI training datasets as a consequence of boilerplate contractual terms. The economic implications are considerable: authors contributing medical and scientific literature receive no royalties from licensing deals that may be valued in the tens of millions of dollars. Cambridge University Press has been identified as a notable exception, having sought input from authors before entering AI

licensing agreements — an approach that, if more widely adopted, would better align commercial licensing practices with the interests of knowledge creators.

From the perspective of an AI developer or research institution seeking to use such content, the practical effect is that the publisher, rather than the author, is the relevant counterparty for any licensing negotiation. However, the legitimacy of the underlying copyright transfer must be assessed on a case-by-case basis: where a contract predates the emergence of AI training as a commercially relevant use, it may be arguable that the transfer did not extend to uses not contemplated at the time of execution, depending on the applicable national contract law.

4.3.3 Machine-Readable Opt-Out Mechanisms and Regulatory Reinforcement

Article 4(3) of the CDSM Directive provides rightsholders with the ability to reserve their rights with respect to TDM for purposes other than scientific research by expressing their objection in a machine-readable manner. This opt-out mechanism has become a key instrument through which publishers restrict the unauthorised use of their content for AI training, and its scope and implementation have significant practical implications for any entity seeking to build training corpora from academic literature.

The standard technical vehicle for implementing opt-outs is the robots.txt file, which instructs automated crawlers and data mining tools to refrain from accessing specified content. Most major academic publishers have incorporated TDM and AI training restrictions into their robots.txt files, as well as into their general terms of service. The Hamburg Regional Court (Landgericht Hamburg) has, in related proceedings, considered whether restrictions expressed in natural language — rather than in a formally structured machine-readable format — also qualify as valid opt-outs under Article 4(3), given that AI systems are technically capable of processing natural language instructions. This question remains unsettled and has significant implications for the interpretation of publisher terms of service as opt-out mechanisms.

The regulatory framework reinforcing these opt-outs has been strengthened by the EU AI Act. Article 53(1)(c) of Regulation (EU) 2024/1689 requires providers of General Purpose AI (GPAI) models to implement a policy ensuring compliance with EU copyright law, explicitly including the opt-out mechanism under Article 4(3) CDSM. This obligation applies irrespective of where the model was trained, thereby extending the reach of EU opt-out protections to non-European AI developers whose models are placed on the EU market.

The GPAI Code of Practice, published in July 2025, further operationalises these obligations. Signatories are required to respect machine-readable opt-outs such as robots.txt, to exclude from their training datasets websites known to host infringing content, to implement technical safeguards against infringing outputs, and to designate a dedicated contact point for copyright complaints. While the Code is not itself legally binding, compliance with it is expected to constitute evidence of due diligence under the AI Act framework.

4.3.4 Open Access Licences and Creative Commons Conditions

Open access (OA) articles published under Creative Commons (CC) licences present a distinct set of licensing constraints. Although OA content is designed to be freely accessible and redistributable, the specific conditions of the applicable CC licence may restrict or complicate its use for LLM training, depending on the nature of the model and the commerciality of the intended application.

The CC-BY licence requires attribution to the original work. In the context of AI training, compliance with the attribution requirement may be fulfilled by linking to the source dataset; however, whether the training process itself triggers the attribution obligation is contingent on whether such use constitutes a "use" requiring copyright permission in the applicable jurisdiction. Where a mandatory TDM exception applies, the use may not require permission at all, and the licence conditions — including attribution — would accordingly not be engaged.

The CC-BY-NC (NonCommercial) licence presents a more substantive constraint. It expressly prohibits commercial use of the licensed works. Training a commercial LLM on NC-licensed medical literature would violate this restriction unless the applicable jurisdiction's TDM exception overrides the contractual condition. This is a critical point of uncertainty: while Article 3 CDSM (scientific research TDM) is a mandatory exception that cannot be contractually overridden, Article 4 CDSM (general TDM) is subject to opt-outs and may not override NC licence conditions where a valid opt-out has been expressed. The CC-BY-ND (NoDerivatives) licence similarly prohibits use as training data if the training process is characterised as the creation of a derivative work — a classification that remains legally contested across jurisdictions.

Creative Commons has issued official guidance stating that, as a matter of copyright law, the use of works to train AI models should be considered non-infringing by default, provided that access to the works was lawful at the time of ingestion. However, this position is contested and does not resolve the question of whether specific licence conditions — particularly NC and ND clauses — may impose contractual restrictions that operate independently of the copyright exception analysis. In jurisdictions where CC licence conditions are enforceable as contracts, a breach of the NC or ND condition could give rise to liability even where no copyright infringement has occurred.

The CC-BY-SA (ShareAlike) licence introduces a further consideration: if AI models trained on SA-licensed content are characterised as adaptations under copyright law, the resulting models would be required to be released under an equivalent licence. This condition is practically incompatible with proprietary commercial AI development and represents a significant legal risk for developers who train on SA-licensed biomedical literature without prior legal assessment.

4.3.5 Subscription Agreements, Database Licences, and the Lawful Access Requirement

Beyond copyright transfer and open-access licence conditions, the use of medical literature for AI training is further constrained by the contractual terms governing access to subscription-based databases and licensed content repositories. Publisher subscription agreements — including those governing access to major platforms such as Scopus, Web of Science, and PubMed's licensed full-text content — typically include express restrictions on automated downloading and data mining beyond the scope of the applicable TDM exceptions. These contractual restrictions operate as an additional layer of control, irrespective of whether the technical act of data extraction would independently constitute copyright infringement.

A foundational legal requirement with direct implications for both TDM exceptions and contractual compliance is the lawful access condition. Both Articles 3 and 4 of the CDSM Directive stipulate that the TDM exception applies only where the entity undertaking mining has lawful access to the works in question. Lawful access is satisfied by accessing content through an institutional subscription, a legitimate open-access repository, or another authorised channel. By contrast, the systematic scraping of paywalled content — whether or not protected by technical access controls — does not satisfy the lawful access requirement and therefore falls

outside the scope of the statutory exceptions. Paywalled content obtained through circumvention of access controls is also exposed to liability under the provisions of the CDSM Directive governing technological protection measures, as well as under national implementations thereof.

4.3.6 The Interaction Between Contractual Restrictions and Mandatory TDM Exceptions

A legally significant distinction governs the enforceability of contractual restrictions against the TDM exceptions under the CDSM Directive. Article 3, which establishes the TDM exception for scientific research purposes, is expressly designated as a mandatory and imperative exception. Any contractual clause that purports to restrict or exclude TDM activities falling within the scope of Article 3 is unenforceable as a matter of EU law. This applies equally to restrictions contained in publisher subscription agreements, database licences, and copyright transfer agreements. The mandatory nature of Article 3 ensures that qualifying research institutions cannot be contractually denied the right to mine lawfully accessed works for non-commercial scientific research purposes.

Article 4, which covers general TDM without a requirement of scientific research purpose, operates on a fundamentally different basis. This exception is subject to the opt-out mechanism under Article 4(3), and rightsholders may restrict its application both through machine-readable mechanisms (such as robots.txt) and through express contractual terms in licensing agreements. Entities whose TDM activities cannot be characterised as scientific research — including commercial AI developers — therefore face a dual layer of restriction: they are not entitled to rely on the mandatory protection of Article 3, and they are fully exposed to both the machine-readable opt-outs and the contractual exclusions that publishers have systematically incorporated into their subscription and licensing terms.

The foregoing analysis demonstrates that the legal framework governing the use of medical scientific literature for LLM training is substantially more complex than a straightforward application of the CDSM TDM exceptions would suggest. Licensing agreements, copyright transfers, open-access licence conditions, machine-readable opt-outs, and the mandatory/non-mandatory distinction between Articles 3 and 4 collectively define a layered compliance environment that requires systematic assessment prior to any large-scale ingestion of academic content for AI training purposes.

4.4 AI and health-sector regulatory considerations

The deployment of AI systems in the healthcare sector, and in particular the use of medical scientific literature as training data for large language models, is subject to a compound regulatory framework that combines the horizontal obligations of the EU AI Act (Regulation 2024/1689) with sector-specific instruments derived from medical device regulation. Understanding this framework requires analysing not only each regime in isolation, but also the points of tension and complementarity between them.

4.4.1 Risk Classification and High-Risk AI in Healthcare

Under the EU AI Act, AI systems intended for use in healthcare contexts are predominantly classified as high-risk pursuant to Annex III of the Regulation. This classification encompasses systems employed as medical devices or in vitro diagnostics, as well as systems used for clinical diagnosis, treatment recommendations, or resource allocation. The high-risk designation

triggers a comprehensive set of obligations, of which Article 10 — governing data and data governance — is among the most consequential for entities training models on medical literature.

4.4.2 Article 10: Data Governance Requirements

Article 10 establishes binding quality criteria for the datasets used in the training, validation, and testing phases of high-risk AI systems. Datasets must be relevant to the intended purpose, sufficiently representative of the operational context, free from errors to the extent technically feasible, and complete with respect to the characteristics relevant to the system's function. Beyond these intrinsic quality requirements, the provision mandates comprehensive documentation of the entire data lifecycle: design choices, collection processes and provenance, preparation and pre-processing operations, underlying methodological assumptions, availability assessments, bias examination procedures, and identification of data gaps. These documentation obligations are substantive rather than formal, requiring genuine evidentiary support rather than pro forma declarations. The data governance requirements under Article 10 enter into force on 2 August 2026 for high-risk AI systems.

A specific provision of practical relevance for medical AI training is Article 10(5), which establishes a conditional exception permitting the processing of special categories of personal data — including health data — for the purpose of detecting and correcting bias in high-risk AI systems. This exception is subject to six cumulative conditions, including pseudonymisation of the data, implementation of technical access controls, deletion of the data following its use, and registration of the processing in the relevant records. The provision is narrowly scoped and does not constitute a general authorisation for broad health data processing; it is limited to bias-related interventions and is operative only where no less intrusive means are available.

4.4.3 GPAI Obligations: Copyright and Transparency (Article 53)

The EU AI Act imposes a distinct set of obligations on providers of general-purpose AI models (GPAI), which are directly relevant to LLMs trained on medical scientific literature. Article 53, whose IP-related provisions entered into force on 2 August 2025, requires GPAI providers to implement a policy ensuring compliance with EU copyright law, including the opt-out mechanism established under Article 4(3) of the Directive on Copyright in the Digital Single Market (CDSM). This obligation applies regardless of the geographic location in which training was conducted, thereby extending the territorial reach of EU copyright rules to non-EU model providers that make their systems available within the Union. Article 53 further requires providers to publish a sufficiently detailed summary of the content used for training, in conformity with a mandatory template issued by the AI Office.

These transparency requirements are complemented by the GPAI Code of Practice, published on 10 July 2025, which — while voluntary in formal terms — constitutes practically essential guidance for providers seeking regulatory certainty. The Code addresses copyright compliance, including the obligation to respect robots.txt directives and machine-readable opt-outs, to exclude sources known for infringement, and to implement safeguards against infringing outputs. On transparency, it introduces a standardised Model Documentation Form covering licensing arrangements, technical specifications, intended use cases, dataset descriptions, and energy consumption data. The AI Office has indicated that, during the initial implementation year through August 2026, enforcement actions will not be initiated against providers that have signed the Code, provided they are demonstrably working towards compliance.

4.4.4 Interplay with Medical Device Regulation and AI Literacy

The intersection of the AI Act with the Medical Device Regulation (MDR, Regulation 2017/745) is addressed in MDCG 2025-6 guidance issued by the European Commission. AI-based medical devices are required to comply simultaneously with both instruments, without the possibility of substituting one regime's requirements for the other. The data documentation standards imposed by Article 10 of the AI Act are treated as complementary to — rather than duplicative of — the clinical evidence requirements under the MDR, meaning that training data governance documentation must satisfy both sets of criteria. MDCG 2025-6 also confirms the availability of the Article 10(5) exception for bias-related processing of health data within the medical device context.

Since 2 February 2025, the EU AI Act additionally mandates AI literacy requirements applicable to all organisations involved in the development, provision, or deployment of AI systems, including healthcare institutions. Hospitals and health systems deploying AI tools — irrespective of whether those tools are classified as high-risk — are required to ensure that their workforce maintains an adequate and demonstrable level of AI literacy. This obligation reflects the legislator's recognition that regulatory compliance cannot be reduced to technical conformity alone, but requires organisational competence across the entities involved in AI deployment.

4.4.5 Tension Between Data Minimisation and Dataset Representativeness

A structural tension of particular significance for healthcare AI training concerns the relationship between the GDPR's data minimisation principle and the representativeness requirements of Article 10 of the AI Act. The GDPR mandates that personal data be collected only to the extent strictly necessary for the specified purpose. The AI Act, conversely, requires that training datasets be sufficiently representative and complete to support the intended function of the system — a criterion that may be incompatible with minimal data collection when the system must operate across diverse patient populations or clinical scenarios. In the domain of medical AI, this tension is especially acute: an insufficiently representative training corpus risks systematic bias and degraded diagnostic performance, while an overly expansive corpus risks processing personal data beyond what is legally permissible. Organisations operating in this space must develop documented justifications that satisfy both regimes simultaneously, typically through the combination of pseudonymisation, purpose limitation, and the specific exemptions available under Article 10(5) of the AI Act and Article 9(2)(j) of the GDPR, the latter of which permits processing for scientific research purposes subject to appropriate safeguards.

5 Intellectual property framework for medical literature

5.1 What content may be protected

The determination of what constitutes protectable subject matter in the context of text and data mining (TDM) of medical scientific literature requires analysis across two distinct but frequently concurrent legal frameworks: copyright as applied to individual works, and the sui generis database right applicable to structured collections of content. These frameworks differ in their doctrinal foundations, conditions of subsistence, and scope of protection, yet they operate in parallel and may apply simultaneously to a single act of access or extraction.

Table 5. Potential rights applicable to different components of medical publications.

Content element	Possible protection	Typical rightsholder	Main reuse concern for AI training
Full-text journal article	Copyright	Usually the publisher, often following copyright transfer by the author	Reproduction and ingestion for training may require reliance on a TDM exception or licence
Abstract	Copyright, depending on originality and length	Usually the publisher or rights holder controlling the publication	May be treated differently in practice, but still requires rights assessment before large-scale reuse
Table	Copyright where the selection or arrangement is original; possibly database-related protection in some contexts	Publisher or other rights holder controlling the published work	Extraction and reuse may raise distinct issues because tables often contain condensed scientific results
Figure, chart, graph, or image	Copyright	Publisher or other visual-content rightsholder	Particularly sensitive because visual elements may be reproduced more directly and may have separate licensing conditions
Supplementary material	Copyright; potentially database right depending on structure and content	Publisher, repository operator, or other rights holder depending on publication model	Often overlooked during ingestion, despite being subject to the same or additional restrictions
Bibliographic metadata	Often limited copyright relevance, but may still be subject to database-related protection or contractual access restrictions	Database provider, platform operator, or publisher	Use may appear low risk, but large-scale extraction can still engage database rights or licence terms

Database record or platform-level corpus	Sui generis database right; contractual restrictions; possibly copyright in database structure	Database maker, platform operator, or publisher	Large-scale scraping or systematic extraction may infringe database rights or breach access terms even if individual records are not strongly protected
Consortium-created dataset derived from collected literature	Potential database right in the resulting dataset structure or arrangement; possible continued constraints from source materials	The consortium partner or entity responsible for dataset creation, subject to upstream rights constraints	Downstream reuse must account both for rights in the new dataset and for continuing obligations linked to the original sources

5.1.1 Literary Works and the Originality Threshold

Under EU copyright law, scientific journal articles qualify as literary works provided they satisfy the originality requirement as articulated by the Court of Justice of the European Union (CJEU) in *Infopaq International A/S v Danske Dagblades Forening* (C-5/08). That standard — the author's own intellectual creation — does not demand artistic merit or significant creative effort; it requires only that the work reflects free and expressive choices made by a human author. The practical consequences of this low threshold are significant: protection extends not solely to complete articles but also to individual paragraphs, structural arrangements, and brief excerpts. The CJEU confirmed in *Infopaq* that the reproduction of as few as eleven consecutive words may constitute an infringing act where those words embody sufficient original expression. This principle is directly and materially applicable to large-scale automated processing of scientific publications, including corpus construction and tokenisation operations underlying LLM training pipelines.

5.1.2 Non-Textual Elements Within Scientific Publications

Protection under copyright law is not confined to the textual body of scientific works. Within a single publication, multiple independently protectable works may coexist, each governed by its own rights configuration. Figures, graphs, photographs, and diagrams that reflect original compositional or design choices attract copyright protection as artistic or literary works in their own right, independently of the article in which they appear. Code samples, supplementary data files, and presentation materials included within or accompanying a publication may similarly constitute separate protectable works. This multiplicative layering of rights within a single artefact substantially increases the rights clearance burden associated with large-scale corpus construction, as each element may require independent authorisation from potentially distinct rights holders.

5.1.3 Elements Excluded from Copyright Protection

Not all content within scientific literature attracts legal protection. Raw empirical facts, isolated data points, and abstract scientific ideas fall outside the scope of copyright, which protects the expression of ideas rather than the ideas themselves. Numerical measurements, chemical parameter values, clinical endpoints, gene sequences presented as factual records, and similar

empirical data carry no inherent copyright protection regardless of the effort or expense involved in their production or collection. This distinction between protectable expression and unprotectable factual content is foundational to any licensing or exception-based analysis. It does not, however, resolve all access questions in isolation: such data is frequently embedded within or derived from an otherwise protected work or a qualifying database, and the means by which it is extracted may independently engage other legal regimes.

5.1.4 Database Copyright Applicable to Curated Collections

Where scientific content is organised into collections that reflect original selection or arrangement criteria — such as structured clinical-trial corpora, systematically curated assay-result repositories, or thematically organised bibliographic databases — the resulting compilation may attract copyright protection as a database work pursuant to Article 3 of Directive 96/9/EC on the legal protection of databases. The protection in such cases attaches to the intellectual choices made in selecting or arranging the contents, not to the individual data items themselves, which remain unprotected in isolation. This layer of protection subsists independently of any copyright vesting in the individual works contained within the database, and is particularly relevant where scientific publishers or data infrastructure providers have exercised editorial judgement in the construction of their repositories.

5.1.5 The Sui Generis Database Right

Directive 96/9/EC also establishes a sui generis database right that operates independently of any originality requirement. This right vests in the maker of a database who can demonstrate substantial investment in obtaining, verifying, or presenting the contents of that database. Unlike copyright, which rewards creative choices, the sui generis right is designed to protect economic investment in the production and maintenance of informational infrastructure, irrespective of the intellectual effort involved in its organisation. Full-text scientific platforms, aggregated citation and abstract indexes, and curated reference databases are strong candidates for protection under this regime, given the material financial and organisational investment typically involved in their construction, quality assurance, and ongoing curation. The right confers on its holder the ability to prevent the extraction or re-utilisation of qualitatively or quantitatively substantial parts of the database contents by third parties.

5.1.6 The Dual-Layer Protection Model and Its Implications for TDM

The legal framework applicable to medical scientific literature is therefore characterised by a dual-layer structure: article-level copyright subsisting in individual works and, simultaneously, a platform-level sui generis right potentially subsisting in the database as a whole. These two regimes operate independently of one another, and any TDM activity engaging with content from a qualifying platform must address both layers separately. Critically, the stacking of these rights means that even where the use of an individual article might be permissible under an applicable exception or licence, large-scale extraction of content from a qualifying database platform may independently constitute an infringement of the sui generis right, regardless of the copyright status of the individual items extracted.

This dual-layer model has been explicitly recognised at the legislative level. The recitals to Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (DSM Directive) confirm that the TDM exceptions provided under Articles 3 and 4 apply regardless of whether the relevant content is protected by copyright alone, the sui generis database right alone, or both regimes concurrently. This cross-cutting scope is designed to ensure that the utility of the TDM exceptions is not defeated by rights-stacking arrangements. Nevertheless, it does

not eliminate the requirement for a careful analysis of which regime or regimes apply in any given operational context, the conditions under which the relevant exception is available, and any limitations or contractual overrides that may affect its application in practice.

5.2 Permitted, restricted, and prohibited uses

The legal framework governing the use of scientific literature in the training of large language models requires a differentiated analysis of the acts involved at each stage of the processing pipeline. For the purposes of this project, the applicable regulatory instruments are the Directive on Copyright in the Digital Single Market (DSM Directive 2019/790), the InfoSoc Directive (2001/29/EC), and the Database Directive (96/9/EC). Within this framework, permitted, restricted, and prohibited uses are determined not by a single legislative rule but by the intersection of the nature of the act performed, the category of entity performing it, the legal status of the content accessed, and whether relevant opt-out or reservation mechanisms have been activated by rightsholders.

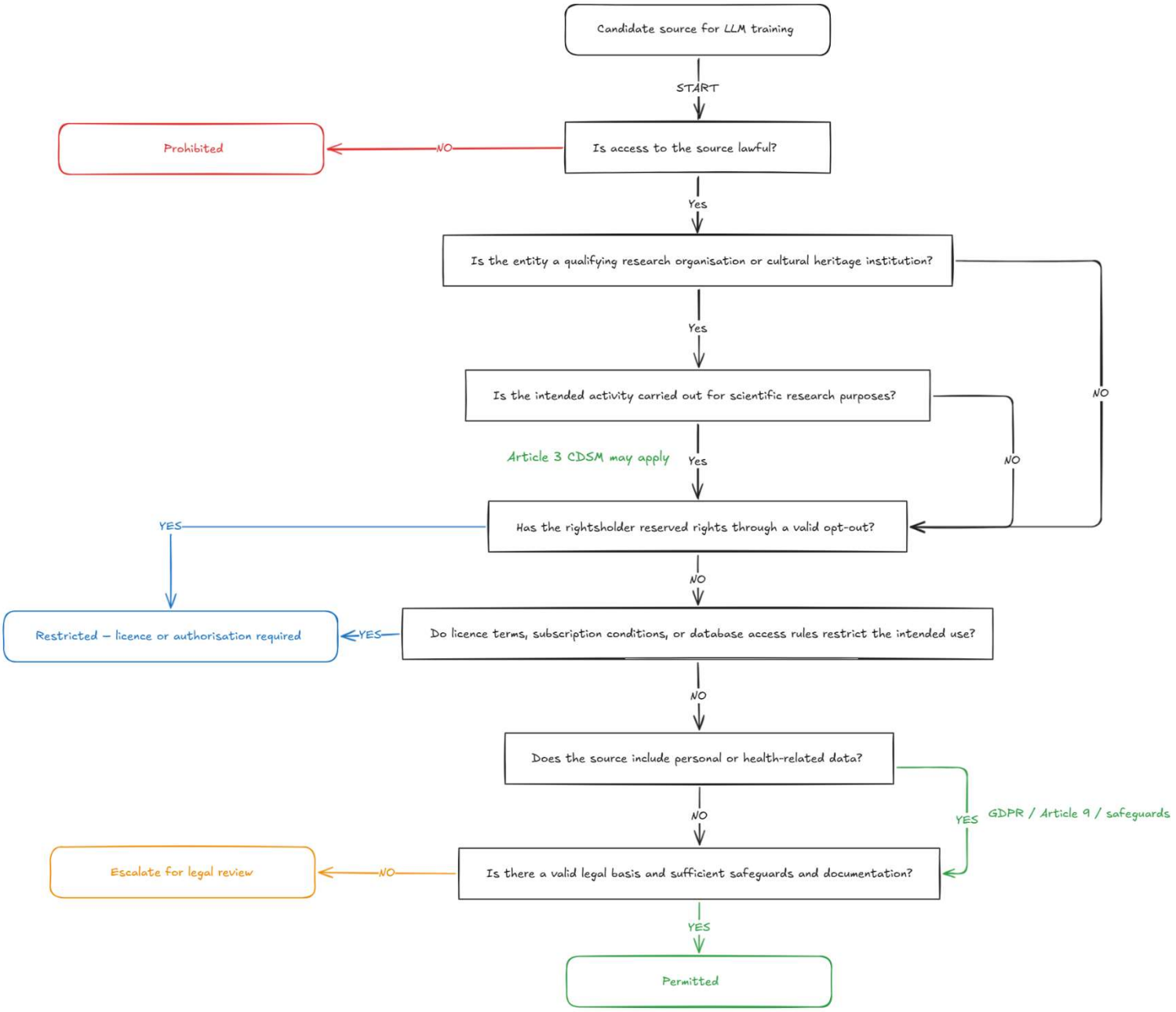

Figure 2. Decision tree for determining whether a source may be used for LLM training.

protocols such as the W3C TDM Reservation Protocol — while for subscription-based content, reservation may be effected through the terms of the applicable licence agreement. A compliant Article 4 workflow therefore requires active, per-source verification that no machine-readable reservation exists, ongoing monitoring of evolving technical standards governing such reservations, and documented evidence of lawful access in respect of each segment of the training corpus.

At the opposite end of the risk spectrum, content that has entered the public domain — including works whose copyright term has expired and government publications that are not eligible for protection — is entirely free of copyright and sui generis database constraints. TDM and model training activities conducted exclusively on such material require no statutory exception, opt-out compliance, or rightholder authorisation. Similarly, individual scientific data points, raw measurements, and factual elements that lack the original creative selection or arrangement required for copyright protection are not copyrightable as such; their use in TDM pipelines represents the lowest-risk category of training data from a copyright standpoint. It should be noted, however, that the absence of copyright protection does not immunise corpus use from all legal constraints: data protection obligations under the General Data Protection Regulation (GDPR), confidentiality duties, trade secrets law, and the terms of service of data platforms may independently restrict access to or use of such material even where copyright has lapsed or never attached.

5.2.2 Restricted and Legally Uncertain Uses

The legal status of large language model training is not fully resolved under the existing DSM exception framework. LLM training involves a sequence of legally relevant acts that extends materially beyond the initial extraction phase of TDM: it encompasses iterative reproduction of corpus data during model optimisation, the creation and retention of persistent intermediate files, and the long-term storage of training corpora. Whether these downstream acts fall within the scope of the reproduction and extraction authorised by Articles 3 and 4 DSM — provisions that are expressly framed around reproductions performed for the purpose of analysis — remains a matter of substantive legal uncertainty. EU legal scholarship has identified a structural gap between the TDM exception framework and the broader technical architecture of LLM training, noting that multiple inference and fine-tuning passes may constitute independent acts of reproduction not fully addressed by the existing exceptions. This uncertainty is directly relevant to the present project and warrants continuous monitoring as regulatory guidance and judicial interpretation develop.

In this connection, the pending CJEU preliminary reference in Like Company v Google Ireland (Case C-250/25) represents the first direct judicial test of whether LLM training on in-copyright content constitutes reproduction within the meaning of the InfoSoc Directive and whether such training may qualify under the Article 4 DSM exception. The outcome of this case will materially clarify the legal boundary between TDM and AI training for EU practitioners and will inform the risk assessment of projects of the nature described in the present report.

The applicability of the temporary and transient copies exception under Article 5(1) of the InfoSoc Directive also warrants specific consideration and, on analysis, exclusion as a viable authorisation basis for corpus creation. That provision excepts copies that are transient or incidental, that form an integral and essential part of a technological process, that possess no

independent economic significance, and that serve solely to enable a lawful network transmission or a lawful use of the work — all conditions being cumulative. Large-scale training corpora, persistent pre-processing files, and tokenised datasets stored for ML or LLM training do not satisfy these requirements: they are purposefully created, they are economically significant as direct training inputs, and they are retained for extended periods well beyond any transient transmission event. The CJEU confirmed this position in Infopaq I and Infopaq II, establishing that even brief automated copies may engage the reproduction right under Article 2 InfoSoc and thereby reinforcing that the temporary-copy exception cannot serve as a general safe harbour for corpus creation in the absence of a more specific TDM exception or explicit authorisation.

5.2.3 Prohibited Uses

Notwithstanding the authorisations that Articles 3 and 4 DSM confer in respect of TDM acts, neither provision extends to the further distribution of the mined works, or of substantial portions of protected databases, to third parties. The redistribution of raw full-text corpora — including scraped PDF collections or compiled article repositories — constitutes a separate act of communication or making available that falls outside the scope of the TDM exceptions and exposes the distributing party to liability under both copyright law and the sui generis database right. Compliant data-sharing strategies must therefore be confined to the transfer of derived artefacts — such as statistical outputs, embedding vectors, or model weights — from which the protected expression of the underlying works cannot be directly extracted or reconstructed.

A further category of prohibited conduct concerns the generation of model outputs that reproduce substantial or near-verbatim portions of protected works. The legality of training-stage activities and the legality of output-stage activities are analytically independent: even where corpus assembly and model training have been lawfully conducted under a TDM exception or applicable licence, a model that generates outputs reproducing recognisable passages from protected works may separately infringe the reproduction right or the right of communication to the public under Article 3 of the InfoSoc Directive. This risk is particularly acute where the training corpus includes paywalled or subscription-access scientific articles, since prompted generation of verbatim or near-verbatim passages could constitute an unlicensed making available of the underlying work to end users. Technical risk-mitigation measures — including training-data deduplication, memorisation testing, output regurgitation filters, and end-user policies restricting verbatim retrieval of specific articles — function both as compliance safeguards and as potentially significant factors in any future proportionality assessment by courts or regulatory authorities.

5.3 Open access versus restricted-access literature

The legal permissibility of using scientific literature for text and data mining (TDM) and large language model (LLM) training is fundamentally conditioned by the access regime under which that literature is made available. A binary distinction between open-access (OA) and restricted-access or subscription-based literature is, however, an oversimplification: within each category, the operative legal constraints vary substantially depending on the specific licence terms attached to each corpus and the commercial or non-commercial character of the training activity in question. This section analyses the principal legal configurations arising under OA and

subscription frameworks, with reference to the applicable provisions of Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (DSM Directive).

5.3.1 Open-Access Literature under Permissive Licences (CC BY)

Scientific publications released under a Creative Commons Attribution (CC BY) licence represent the most permissive end of the OA spectrum. The CC BY licence grants broad rights of reuse, reproduction, adaptation, and redistribution, subject only to the obligation to provide appropriate attribution and, where applicable, to indicate that modifications have been made to the original work. For TDM and LLM training purposes, this licence configuration renders the use of the corpus permissible as a matter of contractual right, without the need to invoke any statutory TDM exception under national or EU law.

This distinction is legally significant. Where the operative basis for TDM is a contractual licence grant rather than a statutory exception, the legal permissibility extends globally and is not bounded by the beneficiary restrictions, purpose limitations, or territorial scope that characterise the EU's Article 3 and Article 4 TDM exceptions under the DSM Directive. Provided that attribution obligations are duly observed and any required indication of modifications is included in the training pipeline documentation or output metadata, the use of CC BY corpora for both non-commercial and commercial LLM development is permissible by default. Compliance is therefore licence-driven, not exception-driven, and the risk profile associated with CC BY literature is correspondingly low.

5.3.2 Open-Access Literature under Non-Commercial Licences (CC BY-NC)

A substantially different legal position arises where OA literature is published under a Creative Commons Attribution Non-Commercial (CC BY-NC) licence. The non-commercial condition embedded in CC BY-NC operates as a contractual restriction that excludes any use whose primary or incidental purpose constitutes commercial exploitation. The training of commercially deployed LLMs — that is, models intended for integration into products or services generating direct or indirect revenue — falls squarely within the category of commercial use, regardless of whether the training activity itself is presented as research-oriented or internally facing.

No statutory TDM exception under EU law substitutes for the licence requirement in commercial contexts. Article 4 of the DSM Directive, which provides a general TDM exception, explicitly permits rightholders to reserve their rights against commercial uses by machine-readable opt-out mechanisms; a non-commercial licence condition achieves a comparable restriction through the contractual channel. Accordingly, organisations seeking to use CC BY-NC corpora as part of a commercial LLM training pipeline are required to secure a separate, purpose-specific licence from the relevant rightholders before initiating any training activity. The absence of such a licence exposes the organisation to breach-of-contract claims and, where the non-commercial condition is read as a condition of the licence grant rather than a covenant, to copyright infringement liability.

5.3.3 Interaction Between OA Status, Lawful Access, and DSM Exceptions

Open access publication satisfies the threshold condition of being "lawfully accessible" that is required to invoke either Article 3 or Article 4 of the DSM Directive. This condition is structural:

a work must be accessible to the relevant actor through a channel that does not itself constitute an infringement for the TDM exception to be triggered. OA publications, by definition, meet this requirement. However, OA status does not operate as a blanket authorisation that eliminates all further legal constraints; licence conditions that extend beyond the baseline of copyright protection — including non-commercial clauses, share-alike obligations, or no-derivatives restrictions — remain fully applicable and must be independently assessed.

The legal analysis therefore proceeds in two layers. The first layer concerns whether lawful access exists in order to trigger the relevant TDM exception. The second layer requires an independent assessment of whether the applicable licence conditions are compatible with the intended training activity. Within the scope of Article 3 of the DSM Directive, which applies exclusively to research organisations and cultural heritage institutions acting for non-commercial scientific research purposes, Article 7 of the DSM Directive renders contractual provisions that seek to override or restrict the exception unenforceable. Outside this protected sphere — that is, for all commercial actors and for research activities that do not fall within the Article 3 definition — licence conditions imposed by authors, publishers, or aggregators continue to bind users and cannot be displaced by the statutory exception.

5.3.4 Subscription-Based Literature: Overlapping Rights Regimes and Restrictive TDM Policies

Subscription-based scientific literature — comprising the majority of high-impact publications in clinical and biomedical domains — operates under a dual layer of intellectual property protection that is more restrictive than either the copyright or the database right considered in isolation. At the article level, the author or assignee publisher holds copyright over the text, figures, and data embedded in each publication. At the platform level, the aggregation and systematic organisation of large volumes of scientific articles gives rise to a sui generis database right vested in the database producer, covering the investment made in obtaining, verifying, and presenting the database contents. Large-scale corpus extraction for TDM purposes triggers both layers simultaneously, requiring authorisation under each.

Major scientific publishers have responded to the commercial AI training context by channelling any permitted TDM through tightly controlled API-only access mechanisms. These mechanisms restrict the scope of permitted extraction to non-commercial research purposes and impose a series of operational constraints including rate ceilings, API-key authentication requirements, prohibitions on the retention of local copies beyond the duration of the licensed activity, data-retention ceilings, and output-licensing requirements that may condition the use of derived models. The practical effect of these policies is to narrow the practical availability of subscription corpora for commercial AI development considerably, even where a valid subscription agreement exists.

Direct extraction of subscription literature by means of web-crawling or PDF scraping, where such activity occurs outside authorised access channels, generates concurrent legal exposure across multiple heads of liability: breach of contract under the subscription or site-licence agreement; copyright infringement in respect of each extracted work; infringement of the sui generis database right; and, where technical protection measures are circumvented to facilitate extraction, anti-circumvention liability under Article 6 of Directive 2001/29/EC. The cumulative litigation risk associated with unauthorised extraction from subscription platforms is therefore substantially greater than that arising from the use of OA literature, and organisations engaged

in building medical LLMs should treat this avenue as effectively foreclosed in the absence of express contractual authorisation.

5.3.5 Publisher TDM Licences as Effective Reservation of Rights under Article 4 DSM

The reservation of rights mechanism established by Article 4(3) of the DSM Directive allows rightholders to opt out of the general commercial TDM exception by expressing a machine-readable reservation attached to the content. Publisher TDM licence terms — particularly those conditioning access on non-commercial research use — constitute a functionally equivalent exercise of this opt-out mechanism, whether they take the form of a machine-readable reservation or of express contractual prohibitions within the subscription agreement.

Many subscription licences entered into by major scientific publishers now contain explicit clauses prohibiting the use of licensed content for general-purpose or commercial AI training. These clauses transform what would otherwise be a de facto access restriction into an unambiguous contractual exclusion and, in some formulations, into a basis for intellectual property infringement claims. For LLM developers, the practical consequence is that compliance with subscription licence terms must be verified not only at the point of initial access but also at the point of downstream use, since the permitted scope of access for the purpose of reading or searching is generally narrower than the scope of use required for large-scale corpus extraction.

5.3.6 Research-Organisation TDM under Article 3 DSM: Legal Permissibility and Practical Constraints

For qualifying research organisations — defined under the DSM Directive as universities, research institutes, and other entities whose primary goal is to carry out scientific research or to conduct educational activities including the carrying out of scientific research — Article 3 of the DSM Directive establishes a non-waivable floor of permissibility for TDM activities undertaken within a scientific research purpose. The non-waivable character of this exception, as secured by Article 7 of the DSM Directive, means that contractual provisions seeking to override or restrict it are rendered unenforceable. A research organisation that satisfies both the beneficiary condition and the purpose condition may therefore conduct TDM on lawfully accessible content — including subscription literature accessed through institutional site licences — without obtaining separate authorisation from the rightholder.

In practice, however, the formal legal permission conferred by Article 3 frequently encounters significant structural obstacles. Technical protection measures, API gatekeeping, and platform-level controls that limit the rate and volume of extractable content restrict practical access independently of the legal position. The gap between formal unenforceability of contractual restrictions and practical inaccessibility arising from technical controls is a persistent feature of the subscription-literature landscape and requires coordinated engagement between research teams, institutional legal offices, and publishers.

It is also essential to note that Article 3's double limitation — restricted to a defined beneficiary class and conditioned on a mandatory scientific-research purpose — excludes a broad range of actors that might otherwise consider themselves adjacent to non-commercial research. Commercial R&D entities, journalism organisations, start-up companies, and hybrid public-private consortia fall outside the protected sphere, even where they can demonstrate that their immediate use of the TDM output serves a non-commercial purpose. This restriction has direct

operational implications for projects that involve industry partners or that contemplate any degree of commercial exploitation of the trained model.

5.3.7 Structural Mechanisms to Reduce Legal Uncertainty

The legal landscape described above points toward a set of upstream structural mechanisms that can materially reduce the legal uncertainty associated with TDM and LLM training on scientific literature. The most operationally effective measure is the adoption of open-access policies and funder or institutional mandates that require or incentivise the publication of scientific outputs under permissive licences — in particular CC BY or equivalent frameworks. Where a sufficiently large and representative corpus of medical scientific literature is made available under CC BY, the need to invoke statutory TDM exceptions or to navigate restrictive subscription terms is substantially reduced.

Complementary to OA mandates, TDM-friendly licence frameworks — whether embedded directly in publication agreements, data-sharing agreements between publishers and research consortia, or corpus-access programmes maintained by publishers at the platform level — provide a contractually stable basis for downstream AI development that does not depend on the uncertain and jurisdiction-variable interpretation of statutory exceptions. From a regulatory and policy standpoint, the most coherent approach aligns OA mandates with explicit TDM and training permissions, addressing the gap that arises where OA status satisfies the lawful-access condition but residual licence conditions — such as non-commercial clauses — continue to frustrate permissible commercial or hybrid use cases. Until such alignment is achieved at scale, entities developing medical LLMs are advised to conduct a rigorous two-layer licence audit — verifying both access conditions and use conditions — for every corpus included in their training pipeline.

6 Data source eligibility and compliance criteria

6.1 Types of sources covered

The construction of a legally compliant corpus of medical scientific literature for large language model (LLM) training necessitates a structured typology of source categories, each governed by distinct licensing frameworks, access conditions, and applicable legal instruments. The categories described in this section collectively define the universe of source materials considered within this project, spanning open scholarly infrastructures, bibliographic metadata aggregators, institutional and subject repositories, open-access journals and platforms, full-text aggregators, subscription-based publishers, commercial bibliographic databases, and domain-specific controlled-access infrastructures. For each category, the analysis distinguishes between the permissive, restrictive, or conditional nature of the applicable terms, with particular attention to the requirements of Directive (EU) 2019/790 on copyright in the Digital Single Market (CDSMD) and the EU AI Act, as well as the specific demands of training pipelines intended for commercial deployment.

6.1.1 Open Scholarly Infrastructures and Repositories

Large-scale, openly governed scholarly infrastructures constitute the primary and most legally straightforward category of compliant training data sources. These infrastructures provide structured access to metadata and, in many cases, to full-text content released under public-

domain dedications or open licenses, thereby satisfying the core data governance requirements applicable under the EU AI Act with respect to transparency and provenance documentation. Repositories that adhere to FAIR (Findable, Accessible, Interoperable, Reusable) and TRUST (Transparency, Responsibility, User Focus, Sustainability, Technology) principles are of particular relevance, as they incorporate persistent identifiers, rich and standardised metadata schemas, machine-readable license signals, and robust long-term preservation frameworks. These features are directly material to compliance: persistent identifiers enable reliable provenance tracking across corpus versions; machine-readable license metadata supports automated filtering and segmentation during corpus construction; and stable preservation guarantees that the lawfulness of the source can be verified at any subsequent point in the pipeline.

Where infrastructure-level datasets are released under Creative Commons Zero (CC0) public-domain dedications, their use for AI training purposes is unrestricted across both commercial and non-commercial scenarios, with no attribution obligations imposed on the downstream user. This represents the highest degree of legal certainty achievable in corpus construction and, accordingly, such sources are prioritised wherever their subject coverage is adequate for the domain objectives of the project.

6.1.2 Bibliographic Metadata Sources

Bibliographic metadata aggregators occupying this category provide CC0-licensed datasets covering hundreds of millions of scholarly records, enabling their lawful use for indexing, sampling, corpus profiling, and, where permissible, training without licensing restrictions on the metadata layer itself. Such sources are not sources of full-text content but serve a critical function in license-aware corpus construction: they allow downstream pipelines to identify, classify, and filter works according to their open-access status, license type, and distribution conditions prior to any content ingestion.

Complementary to general bibliographic aggregators, open evidence bases that systematically track the open-access status and Creative Commons license variant of large publication corpora enable automated, license-aware filtering mechanisms to be embedded directly into corpus construction workflows. Similarly, directory-level metadata services that record journal-level licensing policies — including the specific Creative Commons variants under which journals publish — allow programmatic verification of text and data mining (TDM) and AI training permissions at scale before any content retrieval is initiated. These sources collectively form the informational backbone of a compliant data governance pipeline, supporting audit-ready documentation of the licensing basis for each item included in the training corpus.

6.1.3 Institutional and Subject Repositories

Institutional and domain-specific repositories present a considerably more heterogeneous licensing landscape than the open infrastructures described above. Within this category, individual items may be deposited under public-domain dedications, Creative Commons licenses of varying permissiveness, or institutional distribution terms that do not confer third-party reuse rights. This variability requires item-level, rather than repository-level, license verification as a prerequisite for corpus inclusion. Many such repositories expose machine-readable license metadata via standardised interoperability protocols such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) or via RESTful APIs, which facilitates automated license classification at scale. However, where a given item does not carry an explicit permissive license, its inclusion in the training corpus requires either a separate legal analysis establishing the

applicability of a statutory TDM exception under Article 3 or Article 4 CDSMD, or affirmative exclusion from the corpus where no such exception applies.

A specific risk category within institutional repositories arises in connection with preprint platforms that operate under distribution-only licences. Such licences permit the sharing and display of a work but do not authorise reproduction, derivative creation, or use for purposes beyond those expressly specified — including use as training data. The presumption of open reusability that may informally attach to preprint availability is legally unfounded in these cases, and heightened due diligence protocols are accordingly applied to this subcategory during corpus construction.

6.1.4 Open-Access Journals and Platforms

Open-access journals indexed under Creative Commons Attribution (CC BY) licenses, or under functionally equivalent terms, are generally suitable as training data sources for both commercial and non-commercial AI applications. The primary obligations attaching to CC BY use — attribution to the original author and source — are addressable at the corpus documentation level without imposing downstream constraints on the model outputs produced. Where share-alike conditions are incorporated in the applicable license variant (e.g., CC BY-SA), additional analysis is required to determine whether the generation of model weights or inference outputs from the corpus constitutes a derivative work for the purposes of the license and thereby triggers reciprocal licensing obligations.

By contrast, journals operating under Creative Commons Non-Commercial (CC BY-NC) or Non-Commercial No-Derivatives (CC BY-NC-ND) licenses impose material restrictions that render them unsuitable as training data for commercial AI models, unless the intended deployment context falls unambiguously within the non-commercial scope of the applicable license. Given the commercial nature of the AI system contemplated in this project, content published under NC-type licenses is systematically excluded from the training corpus unless a statutory exception independently authorises its use. Directory services that index journal-level licensing conditions at scale serve as the primary filter mechanism for identifying permissible open-access sources and segregating them from restricted-use variants prior to content retrieval.

6.1.5 Open-Access Aggregators

Bulk open-access repositories providing full-text scientific literature under explicitly labelled Creative Commons licenses represent a high-value source category for AI training corpora, particularly where such aggregators segment their holdings by commercial and non-commercial licensing permissions and expose standardised license metadata alongside each record. Aggregators designed specifically to support secondary analysis and large-scale reuse provide not only content but also the provenance and license documentation necessary for downstream compliance verification — a feature of direct relevance to the transparency and data governance obligations imposed by the EU AI Act on providers of general-purpose AI models.

The commercial/non-commercial license split that characterises the holdings of such aggregators introduces a critical design requirement at the corpus construction level: the data ingestion and segmentation logic must enforce use-case boundaries rigorously, ensuring that non-commercial-only subsets are systematically identified and excluded from training pipelines intended for commercial deployment. Failure to enforce this boundary constitutes a license violation that may expose the project to copyright infringement liability and jeopardise the legal basis for commercial exploitation of the resulting model.

6.1.6 Subscription Publishers

Major subscription publishers constitute the category with the most heterogeneous and, in most cases, the most restrictive conditions governing TDM and AI training activities. The principal legal instrument through which these publishers exercise control over their content is the machine-readable opt-out mechanism provided under Article 4(3) CDSMD, which permits rightsholders to reserve all rights for purposes of AI system development, including model training. Where a publisher has deployed a qualifying opt-out — whether through robots.txt directives, API policy declarations, or terms of service provisions meeting the machine-readability standard — reliance on the general TDM exception under Article 4 is legally foreclosed for their content.

Where TDM API agreements have been made available by subscription publishers, such agreements typically constrain use to non-commercial analytical purposes and explicitly prohibit the use of full text or derived outputs for AI model training, unless a specific and separately negotiated contractual authorisation for training use cases has been obtained. Open-access content published by subscription publishers under permissive Creative Commons licenses — most commonly CC BY — remains usable under the terms of the applicable license irrespective of the publisher's general policy stance on TDM, since the license grant is a binding legal instrument that cannot be unilaterally retracted by means of an opt-out notice. The vast majority of paywalled content from this category is, however, effectively excluded from the training corpus absent a specific AI training agreement that expressly permits such use.

6.1.7 Commercial Bibliographic Databases

Commercial bibliographic databases that provide access to full-text scientific literature typically restrict automated harvesting, bulk download, and any form of reuse for AI training purposes through click-through licence agreements and platform terms of service. These instruments ordinarily limit permissible automated processing to non-commercial research activities and prohibit the redistribution of full-text content, derived outputs, or model artefacts generated from database content. Training on full-text content from commercial databases accordingly requires, as a minimum, the existence of a specific contractual right that expressly permits AI training use, or — where no such right has been granted and no applicable opt-out has been activated — reliance on a statutory TDM exception under national or EU law.

A particularly important consideration applicable to this category is the potential for a TDM access agreement to be granted for analytical purposes while AI training remains a specifically carved-out and prohibited use. The scope of permissible activities under any such agreement must therefore be reviewed with precision prior to corpus inclusion, as a TDM licence that is silent on AI training should not be interpreted as authorising it, particularly in light of the increasingly express treatment of AI training as a distinct and separately licensable activity in publishing industry practice. Corpus construction procedures for this project accordingly require affirmative documentation of the legal basis for each item sourced from commercial database environments.

6.1.8 Domain-Specific Controlled-Access Infrastructures

In specialised domains such as genomics and clinical research, the legal framework governing the use of scientific literature for AI training cannot be addressed exclusively through copyright and licensing analysis. Such domains are characterised by dedicated data governance infrastructures that layer consent-based and data-use-restriction-aware access mechanisms on top of, or in lieu of, conventional publication licensing. These infrastructures manage the intersection between publication-level rights and the rights and obligations attaching to the

underlying primary datasets — clinical records, biological samples, genomic sequences — from which the published findings derive. Access to controlled datasets within such infrastructures is conditioned on data access applications, researcher identity verification, and binding data use agreements (DUAs) that specify the scope of permissible downstream processing, including any restrictions on AI training or model development.

Integration of domain-specific controlled-access frameworks into the training data pipeline is essential wherever publication corpora are linked to, or derived from, sensitive primary data. In these scenarios, compliance requires simultaneous adherence to copyright obligations governing the published work and to data protection requirements — including those arising under Regulation (EU) 2016/679 (GDPR) and, where applicable, Regulation (EU) 2024/1689 (the AI Act) — governing the underlying controlled-access datasets. The failure to address both layers of obligation independently represents a significant legal risk for any AI development project operating in the biomedical or clinical research domain and is accordingly treated as a structurally distinct compliance requirement within the framework established by this document.

6.2 Acceptance and rejection criteria

The construction of training corpora for large language model (LLM) development from scientific literature requires the application of rigorous, multidimensional criteria to determine the suitability of each prospective data source. These criteria span two principal domains: legal and licensing compliance, and data quality and representativeness. Both domains are operationally interdependent and must be satisfied concurrently; a source that presents adequate quality characteristics but lacks a valid legal basis for ingestion is inadmissible, and conversely, a legally permissive source that fails quality standards cannot be incorporated without compromising the integrity of the training corpus. The following subsections define the acceptance and rejection thresholds applicable to each domain.

Table 6. Acceptance and rejection criteria for candidate data sources

Criterion	Accept	Reject	Conditional / escalate
Lawful access	The source is accessed through an authorised channel, such as an institutional subscription, licensed platform, or legitimate open-access repository	The source has been obtained through unauthorised scraping, circumvention of access restrictions, or any other non-authorised channel	Access route is unclear and requires verification before ingestion
Copyright and TDM status	The intended use is clearly covered by Article 3 CDSM, or by Article 4 without valid opt-out, or by an applicable licence	The intended use falls outside the relevant exception and no licence or authorisation is available	The applicability of the exception is uncertain and requires legal review

Rightsholder opt-out	No valid opt-out has been identified for the intended use	A valid Article 4(3) opt-out has been identified and prevents reliance on the general TDM exception	The existence, form, or legal effect of the opt-out is unclear
Licensing and contractual terms	Licence terms are compatible with the intended use, including automated processing where relevant	Subscription, platform, or publisher terms prohibit the intended use	Terms are ambiguous, incomplete, or require contractual clarification
Personal or health-related data	No personal data are present, or a valid legal basis and safeguards are clearly identified	Personal or health-related data are present without a valid legal basis or without appropriate safeguards	Personal data may be present and the legal basis or safeguards need further assessment
Source provenance and traceability	Source origin, access route, and rights status can be documented and recorded	Source provenance cannot be reliably established or documented	Provenance is partially known but documentation is insufficient
Open-access licence conditions	The applicable licence is compatible with the intended research or training use	The source is subject to licence terms that clearly prohibit the intended use, such as relevant non-commercial or other restrictive conditions	Licence terms require interpretation in light of the intended activity
Database and platform-level restrictions	No database-right or platform-level extraction issue has been identified for the intended use	Systematic extraction would infringe database rights or breach database access conditions	The level of extraction or scope of reuse requires additional assessment
Strategic suitability for the project corpus	The source is legally usable and operationally appropriate for inclusion in the corpus	The source is legally blocked or unsuitable for compliant integration	The source may be valuable but should only be used after escalation and documented approval

6.2.1 Legal and Licensing Criteria

Legal and licensing criteria constitute the primary gate through which any prospective data source must pass before quality considerations are applied. Admission of a source into the

training corpus without a valid and documented legal basis exposes the data controller to liability under copyright law, contractual law, and, where personal data is involved, data protection law. The following criteria define the minimum legal threshold for acceptance.

Existence of a Licence or Statutory Exception Authorising Text and Data Mining and AI Training

A data source is acceptable only where its governing licence or an applicable statutory exception explicitly authorises both text and data mining (TDM) operations and the downstream use of extracted content for the purpose of AI model training. The mere absence of an explicit prohibition does not constitute permission; positive authorisation is required. Where no contractual licence exists, reliance on the statutory TDM exceptions established under Articles 3 and 4 of Directive (EU) 2019/790 on copyright in the Digital Single Market (CDSMD) is permissible, provided that no valid rights reservation has been exercised by the rightsholder in accordance with Article 4(3) of that Directive. Sources lacking either a permissive contractual licence or an applicable unencumbered statutory basis must be excluded.

Preferred Licence Categories: CC0 and CC BY

Creative Commons CC0 (public domain dedication) and CC BY (attribution-only) licences represent the preferred legal baseline for training corpus construction in both research and commercial deployment contexts. CC0 eliminates all copyright and related rights, conferring unrestricted rights to use, reproduce, adapt, and redistribute content without any conditions. CC BY grants equivalent freedoms subject solely to the obligation to attribute the original author and source in a manner specified by the licensor. Sources governed by either of these licences afford maximum legal certainty and should be prioritised in source selection protocols and ingestion pipeline design. Their use minimises the risk of downstream licence incompatibility in model distribution arrangements.

Case-by-Case Assessment of CC BY-SA Compatibility

Sources licensed under CC BY-SA impose a share-alike (copyleft) obligation requiring that any derivative work be distributed under terms identical to those of the source licence. The applicability of this obligation to trained model weights and model outputs is legally unsettled; however, to the extent that a trained model may be considered a derivative work of its training data, the share-alike condition may render CC BY-SA sources incompatible with proprietary or closed-source model distribution arrangements. Accordingly, CC BY-SA sources must not be treated as functionally equivalent to CC BY without explicit legal review. Acceptance of CC BY-SA sources requires individual compatibility assessment against the intended model deployment model, and the outcome of that assessment must be documented in the data governance record.

Rejection of CC BY-NC and CC BY-NC-ND Sources for Commercial Applications

Licences incorporating a non-commercial restriction — specifically CC BY-NC and CC BY-NC-ND — are categorically unsuitable for training corpora intended for use in commercial AI systems. Both licence types prohibit any form of commercial exploitation, including the development of commercially deployed models trained on such content. The CC BY-NC-ND variant additionally prohibits the creation of derivative works, thereby foreclosing any adaptation or transformation of the source material. Retention of CC BY-NC or CC BY-NC-ND sources within the corpus may be

permissible exclusively for non-commercial research projects, and only where documented operational controls are in place to prevent any commercial downstream application of the resulting model or its outputs. In the absence of such controls, these sources must be excluded.

Absence of a Valid TDM or AI Training Rights Reservation

Content in respect of which a rightsholder has exercised a machine-readable opt-out under Article 4(3) CDSMD, or has contractually reserved rights with respect to AI training through published policy instruments, must be excluded from training corpora by default. Prominent examples of such reservations include the policies published by major scientific publishers — such as Springer Nature's declared reservation of AI training rights — which operate independently of the underlying open access licence status of the content. Ingestion pipelines must incorporate automated mechanisms for the detection and enforcement of these reservations at the point of data collection. Furthermore, commercial publisher licence terms must be reviewed to distinguish between TDM agreements that authorise analytical mining of content for non-generative purposes and those that explicitly prohibit the use of content or derived outputs for model training; these two categories are legally distinct and not interchangeable.

Lawful Access Through Authorised Access Channels

Data ingestion is lawful only where access to the source is established through a recognised legal pathway. Accepted access mechanisms include institutional subscription agreements covering TDM use cases, publisher-authorised API access, and openly accessible repositories. Content obtained through mechanisms not falling within one of these categories does not benefit from a valid legal basis for processing, regardless of the content's downstream licence status. It must be noted that the existence of a permissive content licence does not cure a defect in the legality of the access mechanism through which the content was obtained; the lawfulness of access and the lawfulness of use are legally independent conditions, both of which must be satisfied.

Compliance with Terms of Service and robots.txt Directives

Terms of service governing automated access to web-based content sources, and robots.txt directives specifying permitted crawler behaviour, constitute binding operational constraints on data collection activities. Collection pipelines must incorporate mechanisms to parse and enforce these signals prior to initiating any harvesting activity; compliance cannot be deferred to a post-collection audit. Non-compliance with platform terms of use may give rise to contractual liability independently of, and in addition to, any exposure under copyright or data protection law. Sources for which automated compliance enforcement cannot be technically implemented must be treated as inaccessible pending individual legal review.

Detection and Enforcement of Article 4 CDSMD Opt-Out Signals

Ingestion pipelines must implement automated detection of machine-readable TDM opt-out signals published by rightsholders in accordance with Article 4(3) CDSMD. Such signals may be communicated through multiple channels, including Crossref TDM services, publisher-maintained API metadata fields, and website-level policy declarations. Content bearing a valid Article 4(3) opt-out must be identified and excluded at the point of collection; retrospective

exclusion from the training set after the content has already been processed is legally insufficient, as the processing event itself may already constitute a rights infringement.

Data Protection Compliance for Sources Containing Identifiable Personal Data

Scientific publications containing identifiable personal data — including clinical case reports, qualitative research transcripts, and individual-level datasets — require a valid legal basis for processing under Regulation (EU) 2016/679 (GDPR) prior to ingestion. The scientific or technical purpose of the AI training activity does not automatically supply that basis. Where reliance on the GDPR Article 89 research exemption is considered, the associated safeguards — including pseudonymisation, data minimisation, access controls, and the implementation of appropriate technical and organisational measures — must be documented and operationally in place prior to ingestion, not applied retrospectively. Sources containing personal data for which a compliant processing basis cannot be established must be excluded.

6.2.2 Data Quality and Representativeness Criteria

Beyond legal admissibility, each source must satisfy substantive quality and representativeness thresholds. These requirements derive from both the technical needs of the training process and the compliance obligations imposed by the EU AI Act on the data governance of high-risk AI systems. Quality criteria apply throughout the corpus lifecycle, from initial source selection through ongoing maintenance and documentation.

Relevance to the Intended Domain and Use Case

Each source must demonstrate substantive thematic and methodological relevance to the target domain and intended application context of the AI system. Generic or tangentially related corpora do not satisfy the suitability requirement imposed by Article 10 of Regulation (EU) 2024/1689 (EU AI Act), which requires that training data be relevant and sufficiently representative in relation to the intended purpose of the system. Relevance assessment must be formalised and documented as part of the data governance record, so as to support conformity assessment by competent authorities and to substantiate claims made in technical documentation accompanying the system.

Statistical Representativeness, Completeness, and Error Minimisation

Training corpora must be evaluated for statistical representativeness across the populations, subject areas, and methodological traditions within the intended scope of the AI system, as mandated by EU AI Act Article 10. Gaps in coverage constitute a documented quality deficiency that must be recorded and, where feasible, remediated. Quality assessment processes aligned with the ISO/IEC 5259 series on data quality for analytics and machine learning must be applied throughout the corpus lifecycle to detect, quantify, and remediate incompleteness, internal inconsistency, and unacceptable error rates. These processes must be documented to a level of granularity sufficient to allow independent verification.

Systematic Bias Detection and Documented Mitigation

Systematic bias analyses must examine corpus coverage across multiple dimensions, including language, geographic origin, institutional affiliation, funding source, journal tier, and research topic distribution, in accordance with the bias examination obligations of EU AI Act Article 10. Identified biases must be documented with specificity, including quantification of coverage gaps and their potential effect on model behaviour. Where technically feasible, mitigation strategies — such as targeted corpus supplementation, source reweighting, or stratified sampling — must be implemented and their effect on bias reduction assessed and recorded. Residual biases that cannot be fully remediated must be disclosed in dataset documentation and reflected in system-level documentation.

Repository Trustworthiness: FAIR and TRUST Alignment

Repository selection must prioritise infrastructures that implement FAIR principles, ensuring that content is Findable via persistent identifiers (such as DOI, ORCID, or ROR), Accessible under clearly specified and machine-readable conditions, Interoperable through standard metadata schemas, and Reusable under explicit licence declarations. Repositories must additionally satisfy the TRUST principles — Transparency, Responsibility, User Focus, Sustainability, and Technology — to guarantee long-term preservation integrity, governance accountability, and stable access conditions. Repositories that lack machine-readable licence metadata or provenance information should be avoided, as their inclusion introduces material compliance and auditability risks that cannot be adequately managed through downstream controls.

Explicit Provenance and Curation Documentation

Preferred sources are those accompanied by formal dataset documentation artefacts that describe collection methodology, filtering criteria, de-duplication logic, applied quality controls, and known limitations. Recognised documentation formats include Datasheets for Datasets, Data Cards, Dataset Nutrition Labels, and Data Portraits. Corpora lacking explicit provenance documentation introduce material auditability risks that are structurally incompatible with the traceability obligations of EU AI Act Article 10, and render compliance verification by competent authorities significantly more difficult in practice. The absence of such documentation should be treated as a quality deficiency requiring remediation prior to acceptance.

Exclusion of Synthetic, Contaminated, and Low-Quality Content

Content generated by LLMs or other generative AI systems must be excluded from training corpora to prevent model contamination, the circular reinforcement of training artefacts, and the progressive degradation of output quality through recursive synthetic data ingestion. This standard mirrors the curation approach adopted by recent openly licensed corpora specifically designed for responsible LLM training. Corpus curation pipelines must incorporate explicit detection and exclusion filters targeting synthetic content, retracted publications, expressions of concern flagged by publishers or indexing services, and any other content of demonstrably low quality, irrespective of the licence status of the source. The application of these filters must be logged and the rejection criteria applied must be documented.

Technical Feasibility of Compliant Automated Ingestion

Sources must be accessible through technically implementable, licence-aware ingestion mechanisms capable of enforcing filtering rules — including rights reservation checks, licence classification, and provenance verification — at the point of collection. This requirement is satisfied where sources provide integration with metadata-rich APIs that expose licence type, rights reservation status, and provenance fields. Relevant infrastructure includes OpenAlex, Unpaywall, PubMed Central Open Access, and the Directory of Open Access Journals (DOAJ). Sources for which compliant automated access cannot be technically implemented — for example, due to the absence of machine-readable licensing signals or the lack of an authorised API endpoint — should be excluded from the corpus pending individual legal review and manual clearance procedures. Technical feasibility of compliant ingestion is a necessary, not merely desirable, condition for source acceptance.

6.3 Documentation and traceability requirements

The construction of a training corpus derived from scientific literature creates a set of documentation and traceability obligations that span the entire data lifecycle, from initial source identification to the maintenance of versioned provenance records post-deployment. These obligations arise from multiple converging legal and technical frameworks: the data governance requirements of the EU AI Act, particularly Article 10 on data and data governance for high-risk AI systems; the transparency expectations applicable to general-purpose AI models under the Act's Title VIII provisions; the rights-management regime established by the DSM Directive with respect to text and data mining (TDM); and the accountability obligations derived from the GDPR where the corpus contains personal data. Beyond strict compliance, rigorous documentation constitutes a methodological prerequisite for scientific reproducibility, audit-readiness, and the defensible governance of intellectual property. The following subsections set out the specific documentation requirements applicable across each functional layer of the data pipeline.

6.3.1 Source Registry

The foundational instrument of traceability is a structured source registry that enumerates every data origin contributing to the training corpus. Each source — whether an open scholarly infrastructure, a publisher portal, an open-access aggregation service, or a subscription-based database — must be assigned a stable internal identifier and documented individually within the governance record. The registry must characterise the infrastructural type of each source in terms that carry legal relevance: open scholarly infrastructures such as OpenAlex or Unpaywall operate under open metadata licences and present a different rights profile from open-access aggregators such as the PubMed Central Open Access Subset or Directory of Open Access Journals-indexed publications, which in turn differ from commercial or subscription-based sources accessed under contractual arrangements.

For each registered source, the specific endpoint or access mechanism used for data retrieval must be captured with precision. This includes bulk snapshot download URLs, API base paths, OAI-PMH harvest interfaces, and the version or release identifier of any dataset snapshot used. The access modality — whether authenticated API, anonymous bulk dump, or institutional subscription API — must be recorded alongside the temporal parameters of the access. Where data access is governed by a formal agreement, such as a publisher TDM licence, an institutional

subscription contract, or a data transfer agreement, the relevant contract reference must be linked directly to the source registry entry. For open infrastructure sources accessed without contractual formality, the applicable terms-of-use document and its retrieval date must be referenced to establish the legal basis operative at the moment of collection.

6.3.2 Licensing and Rights Matrix

A licensing and rights matrix must be maintained as an integral component of the source registry, assigning each source a permission classification that specifies whether TDM and AI model training are explicitly authorised, conditionally permitted, or prohibited under the applicable licence terms. This classification must resolve the critical commercial/non-commercial distinction at a granular level: the difference between a Creative Commons Attribution licence (CC BY) and a Creative Commons Attribution-NonCommercial licence (CC BY-NC) has direct consequences for the permissibility of training a model intended for commercial deployment. Similarly, proprietary TDM API terms may authorise certain analytical operations while explicitly excluding use of content for training generative or predictive models, a distinction that must be captured in the rights matrix rather than assumed from the general openness of a source.

The matrix must separately record any opt-outs exercised by rightsholders pursuant to the mechanism established under Article 4 of the DSM Directive. That provision conditions the broad TDM exception on the absence of a rights reservation expressed through machine-readable means, and where such a reservation has been communicated — whether through Crossref TDM service metadata, publisher-level policy headers, or explicit contractual clauses — the affected source must be flagged as ineligible for training under the statutory exception and excluded from the corpus unless an alternative legal basis applies. Publisher-level AI training reservations must be distinguished from general TDM opt-outs, since some licence instruments permit analytical mining for non-generative research purposes while explicitly prohibiting the use of content or derived outputs for model training. Finally, the legal framework relied upon for each source — whether the DSM Directive Article 3 research exception, the Article 4 exception, a negotiated contractual licence, or a combination — must be explicitly documented, including the applicable national jurisdiction and any relevant deviations introduced through domestic transposition of the Directive.

6.3.3 Collection Pipeline Description

Full auditability of the collection process requires that all technical components used for data retrieval be identified with sufficient specificity to enable independent reproduction. This encompasses API client libraries, bulk-download scripts, and harvesting tools, each of which must be documented with version numbers that correspond precisely to the components deployed during the collection run. The pipeline architecture must clarify whether retrieval is performed through structured APIs that expose licence metadata — which is the preferred approach, as it enables automated rights-checking at the point of ingestion — or through unstructured web access, the latter requiring supplementary mechanisms to establish the rights status of retrieved content.

Compliance with site-level access constraints must be documented as part of the pipeline description. The policy applied to robots.txt directives must be stated, as must the crawl-delay and rate-limiting parameters observed during ingestion. Critically, the mechanisms employed to

detect and honour machine-readable TDM opt-out signals must be described as integral components of the pipeline architecture rather than as external considerations. These mechanisms may include the interrogation of Crossref TDM service metadata, the inspection of publisher-provided HTTP policy headers, and cross-referencing against curated opt-out registries maintained by rights management organisations.

The metadata fields captured during ingestion must also be enumerated in the pipeline documentation. At minimum, the record should cover persistent identifiers such as DOIs and PubMed IDs; licence tags and open-access status flags; journal-level ISSNs; and institutional and author identifiers such as ROR and ORCID codes. Where metadata is sourced from interoperable scholarly infrastructure — for example, the OpenAlex CC0 graph or the Unpaywall evidence base — the snapshot date of the metadata source must be captured alongside the content snapshot, since rights information in these systems is subject to change and the state operative at collection time determines the legal basis for inclusion. The licence-based inclusion and exclusion logic applied at the point of ingestion — specifying which licence values or permission flags trigger acceptance or rejection of a document into the training corpus — must be formally recorded as part of the automated filtering layer.

6.3.4 Pre-Processing and Filtering Record

All transformation and filtering operations applied to the raw collected corpus prior to its use in model training must be documented in a dedicated pre-processing record. The de-duplication strategy is of particular importance: the identifier basis used to detect duplicate instances — typically DOI-based canonical de-duplication — must be stated, together with the resolution policy applied where multiple versions of the same work are present, such as a preference for the publisher version of record over a preprint deposit. The choice of version has potential implications both for content quality and for rights compliance, since preprint and published versions may carry distinct licence conditions.

File format inclusion criteria must be documented with justification relative to the intended model use case. The decision to admit only XML full-text versions rather than PDFs, for instance, has direct consequences for the precision of text extraction and the reliability of metadata retention, and such decisions must be recorded transparently. Language detection methods and the scope of languages included in the corpus must be documented explicitly, as language coverage bears upon the representativeness assessment required under EU AI Act Article 10 and on the risk of systematic bias in the trained model's outputs.

The quality control layer applied to the corpus must be specified in terms of the criteria and detection mechanisms used to identify and exclude compromised content. This includes retracted articles, expressions of concern, and duplicate submissions, the detection of which relies on specialised data feeds such as Crossref retraction notifications and PubMed retraction flags. The exclusion of synthetic or machine-generated content from the training corpus must be documented as an explicit policy measure, consistent with emerging best practices in open LLM corpus construction. The presence of model-generated text in training data creates well-documented contamination risks, and the absence of a documented exclusion policy would constitute a material gap in the data governance record.

6.3.5 6.3.5. Provenance and Versioning Metadata

Each source contribution to the corpus must be associated with a precise snapshot date reflecting the state of the external repository at the time of data retrieval. This temporal anchoring is necessary to enable retrospective reconstruction of the legal and factual status of the data at the moment of collection — a requirement that becomes practically significant in the event of a rightsholder challenge or regulatory inquiry. A versioned change log must be maintained for the corpus as a whole, recording all additions, removals, and re-filtering events with timestamps and stated rationale. Removals may be triggered by rightsholder takedown requests, by the subsequent identification of retracted works, or by modifications to the rights matrix following a revision of a source's licence terms; each such event must be individually logged to preserve audit continuity.

Traceability of individual documents to their canonical scholarly identity must be preserved throughout all processing stages through the retention of DOIs or equivalent persistent identifiers. These identifiers must survive the entire pipeline from raw retrieval through tokenisation and, where the architecture permits, be maintained in association with the resulting training tokens to support targeted removal operations. Author identifiers (ORCID) and institutional identifiers (ROR) should be retained in the corpus metadata where available, as they underpin the bias and coverage analysis described in the following subsection and facilitate compliance with any future transparency obligations concerning the representation of specific author communities in training datasets.

6.3.6 6.3.6. Bias and Coverage Assessment Record

The EU AI Act imposes, through Article 10, an obligation on developers of high-risk AI systems to examine training data for biases that could affect system outputs. Compliance with this obligation requires a systematic corpus coverage analysis conducted across multiple dimensions, including research domain distribution, geographic origin of authors and institutions, publication language, funding source affiliation, and journal prestige tier. The analysis must be documented with sufficient methodological detail to permit independent assessment, and the resulting record must form a durable component of the technical file maintained for the system.

Known gaps and imbalances identified through the coverage analysis must be documented alongside the mitigation measures adopted in response — such as the targeted inclusion of underrepresented domain corpora or the application of language-stratified sampling strategies — and residual limitations must be affirmatively disclosed rather than treated as absent by default. For corpora containing health and biomedical literature, domain-specific standards such as the STANDING Together consensus recommendations should be applied to the bias assessment, since these provide methodological criteria calibrated to the particular risks of bias in medical data used for AI applications. The absence of bias documentation cannot be treated as a default assumption of neutrality and constitutes a formal governance limitation that must be reflected in the system's risk profile.

6.3.7 Data Protection Analysis

Scientific literature corpora derived from biomedical and clinical research may contain data that falls within the scope of the GDPR. Patient case reports, qualitative interview transcripts, and publications incorporating individual-level clinical data are categories of content whose inclusion

in a training corpus requires a documented assessment establishing the presence or absence of personally identifiable information and determining whether that information can be considered adequately de-identified for the purposes of further processing. The data protection analysis must be maintained as a formal component of the governance record and must address, at minimum, the categories of personal data present, the applicable GDPR legal basis for processing, and the specific technical and organisational safeguards deployed.

Where personal data is retained in the corpus — whether in full-text content, structured metadata, or both — the safeguards applied must be specified, including pseudonymisation or de-identification procedures, restricted access controls, and data minimisation measures. Where further processing for research purposes is relied upon under Article 89 of the GDPR, a compatibility analysis must be documented demonstrating that the processing is proportionate, that appropriate safeguards are in place, and that the rights of data subjects are not disproportionately affected. For AI systems classified as high-risk under the EU AI Act, data protection compliance documentation must be maintained in conjunction with the data governance record required under Article 10, demonstrating that both regulatory regimes are simultaneously and coherently satisfied.

6.3.8 Documentation Artefacts: Standards and Formats

The foregoing documentation requirements must be embodied in structured artefacts that support both human review and machine-readable audit. Several established frameworks provide appropriate standards for this purpose. Datasheets for Datasets, as originally proposed by Gebru et al. and subsequently refined in the machine learning community, constitute a documentation standard that systematically records dataset motivation, composition, collection process, pre-processing operations, recommended uses, and ethical and legal considerations. Adoption of this format provides a standardised transparency baseline and directly supports the accountability obligations imposed under the EU AI Act with respect to training data governance.

Complementary instruments include Data Cards and Dataset Nutrition Labels, which present provenance, collection methodology, known limitations, and intended use in structured formats suited to communicating data governance posture to downstream model developers, auditors, and regulatory authorities. For corpora intended for use in high-risk or general-purpose AI systems, machine-readable documentation formats acquire additional significance. Open Datasheets encode responsible AI considerations in a form that enables automated risk assessments and supports the regulatory expectation of auditable, authority-accessible training data documentation. Data Portraits serve as corpus-level artefacts that record which specific training data contributed to a large-scale model, enabling downstream users and rightsholders to query the inclusion of particular works. This latter function is of direct relevance to the transparency obligations applicable to general-purpose AI model providers under the EU AI Act, which require the publication of a sufficiently detailed summary of the copyright-protected content used in training, and to the broader accountability framework within which the project operates.

7 Compliance requirements across the data lifecycle

The legal and compliance assessment applicable to the use of medical scientific literature for LLM training cannot be understood as a single verification carried out at the moment of source selection. On the contrary, compliance must be maintained throughout the entire data lifecycle, from the initial identification of candidate sources to the eventual retention review and deletion of the materials or derivative datasets. Each stage of this lifecycle gives rise to distinct legal, organisational, and documentation requirements, and failures at any one of these stages may undermine the lawfulness, traceability, or defensibility of the overall data-processing workflow.

For this reason, the present section adopts a lifecycle-based approach to compliance. It examines the requirements that apply to the collection and acquisition of medical literature, the conditions governing storage and access control, the constraints affecting processing and preparation for AI use, and the obligations associated with data sharing, retention, and deletion. This approach ensures that compliance is embedded as a continuous operational principle rather than treated as a one-off legal review performed at the beginning of the project.

The next figure provides a consolidated overview of this lifecycle-oriented compliance model and illustrates how legal, governance, and documentation obligations accompany each successive stage of the handling of medical literature within the project's training pipeline.

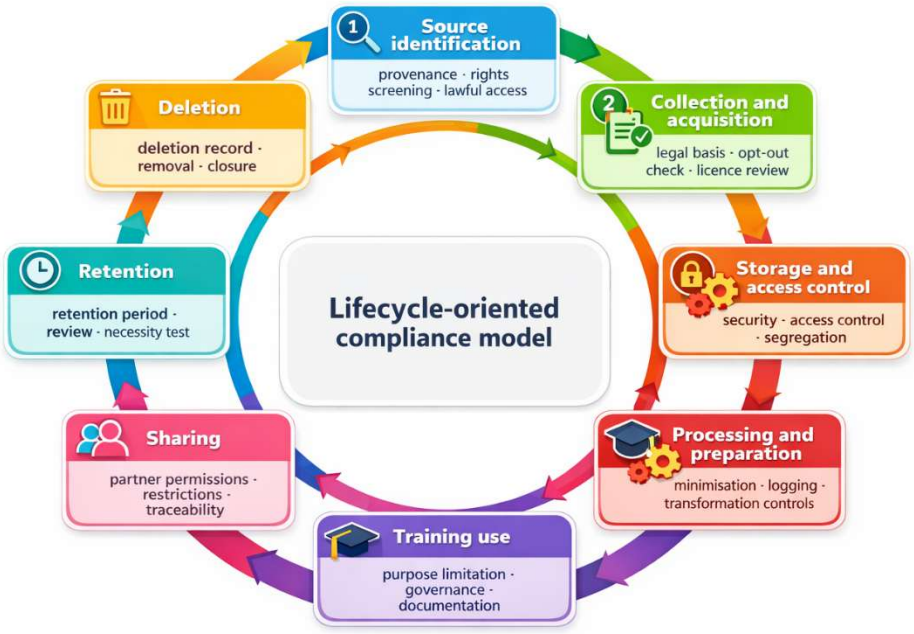

Figure 3. Compliance requirements across the medical literature data lifecycle

7.1 Collection and acquisition

7.1.1 Regulatory Scope and Jurisdictional Framing

The collection and acquisition phase constitutes the legal and operational foundation of the entire training pipeline. Before any data collection activity is initiated, the applicable regulatory perimeter must be formally delineated. At minimum, this entails a comprehensive mapping of obligations arising from the EU AI Act, the Directive on Copyright in the Digital Single Market

(CDSMD, Directive 2019/790/EU, Articles 3–4), and the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679), supplemented by any sector-specific instruments relevant to the target domain — in this case, medical and biomedical scientific literature. Given the cross-border nature of scientific publishing and the international sourcing of training corpora, jurisdictional analysis must precede pipeline design to ensure that the collection methodology achieves compliance across all territories in which the system will be deployed or from which data will be sourced. This preliminary scoping exercise is not merely procedural; it directly conditions the architectural choices made in subsequent pipeline stages.

7.1.2 AI System Classification and Its Implications for Data Governance

Prior to defining collection requirements, the AI system under development must be formally classified under the EU AI Act as either a high-risk system or a general-purpose AI (GPAI) model. This classification is determinative, as each category entails distinct data governance obligations with direct consequences for the collection phase. High-risk systems are subject to the stringent data quality and governance requirements of Article 10 of the EU AI Act, which mandate that training datasets be relevant, representative, free of errors to the extent practicable, and sufficiently complete for the intended purpose. GPAI models, by contrast, are subject to the transparency and copyright-summary requirements introduced by Articles 53 and 55, including the obligation to publish a sufficiently detailed summary of the content used for training. The system classification must be formalised in a scoping decision record at the outset of the project, as it determines the threshold of documentation, auditability, and compliance controls that must be embedded from the collection phase onward.

7.1.3 Source Vetting and Acceptance Criteria

Each candidate repository, publisher, or database must undergo a structured legal and quality assessment against predefined acceptance criteria before being admitted to the ingestion pipeline. Source vetting must cover, at minimum, the following dimensions: licensing conditions and the scope of rights granted; the presence or absence of machine-readable TDM permissions or reservations; data quality indicators applicable to the biomedical domain; and compliance with the FAIR principles (Findability, Accessibility, Interoperability, Reusability) and the TRUST principles governing trustworthy data repositories. Sources that fail any mandatory criterion must be rejected outright or placed in a restricted queue pending further legal analysis. This gatekeeping function ensures that the compliance burden is managed at the point of entry rather than post hoc, and that the provenance of every admitted source is traceable throughout the lifecycle of the training corpus.

7.1.4 Licence-Aware Harvesting Architecture

The harvesting process must be architecturally designed to retrieve and parse machine-readable licence metadata at the article level, enabling automated inclusion or exclusion decisions based on the licence type associated with each individual item. Licence-awareness must be embedded as a functional requirement of the ingestion pipeline itself, not treated as a post-collection filter applied to an already-assembled corpus. This distinction is both technical and legal: if non-compliant content is ingested and subsequently filtered out, the act of initial collection may itself constitute an infringement or a violation of applicable contractual terms. The pipeline must therefore enforce licence-based access control at the point of retrieval, ensuring that content bearing incompatible or ambiguous licence designations is never written to the training corpus.

Priority must be given to sources that expose structured, machine-readable licence information through official APIs or bulk data exports. Examples of infrastructure that satisfies this

requirement include the OpenAlex CC0 metadata graph, the Unpaywall open-access evidence base, the PubMed Central (PMC) Open Access Subset, and the Directory of Open Access Journals (DOAJ) journal-level metadata. These sources enable programmatic licence classification at scale and facilitate reproducible compliance documentation. Access mechanisms that do not expose licence metadata in a structured and queryable format must either be avoided or supplemented with manual legal review, given the compliance risk and the reduction in auditability that they entail.

7.1.5 Licence Compatibility Assessment

Only articles bearing licences that expressly permit text and data mining and AI training for the intended use case — whether commercial or non-commercial — may be included in the corpus. CC0 (public domain dedication) and CC BY licences are generally acceptable for both commercial and non-commercial purposes, as they impose no restrictions on use beyond, in the latter case, attribution. CC BY-SA introduces copyleft obligations requiring that derivative works be distributed under the same or a compatible licence; this condition may be structurally incompatible with proprietary model architectures and must be assessed on a case-by-case basis. CC BY-NC and CC BY-NC-ND licences are restricted to non-commercial contexts and must be excluded from any corpus destined for commercial deployment unless a specific contractual authorisation has been obtained from the rightsholder. Articles operating under restrictive proprietary terms, embargo conditions, or licences that have been superseded by an explicit AI-training reservation must be excluded unconditionally in the absence of such contractual authorisation.

Detection and Enforcement of TDM Opt-Outs and Rights Reservations

The collection pipeline must implement robust mechanisms to detect and honour machine-readable TDM and AI-training opt-outs or rights reservations exercised by rightsholders under Article 4(3) of the CDSMD. This obligation encompasses both publisher-level blanket reservations — such as the Springer Nature reservation policy, which operates at the portfolio level — and item-level signals expressed through dedicated metadata fields or structured protocols. Detection must be performed dynamically at the point of harvesting, as reservation statuses may change over time; content flagged as reserved must be automatically excluded from the corpus and logged with sufficient granularity for subsequent audit and regulatory enquiry. Static or pre-compiled blocklists are insufficient: the pipeline must query reservation signals in real time or at sufficiently frequent intervals to ensure that changes in reservation status are captured before ingestion proceeds.

In operational terms, integration with Crossref's TDM infrastructure and with publisher-specific TDM APIs is required to retrieve authoritative and up-to-date access rights and opt-out signals associated with individual Digital Object Identifiers (DOIs). Publisher-provided machine-readable policy signals — including robots.txt directives, API licence flags, and dedicated metadata fields — must be consumed and acted upon as part of the rights-resolution workflow that precedes any download. This integration ensures that the pipeline's compliance posture reflects the current state of rightsholder intentions rather than historical metadata snapshots.

7.1.6 Pre-Download Rights Resolution

A formal rights-resolution step must be executed for each candidate article prior to downloading its full text. This step involves confirming that the applicable licence, the absence of any TDM or AI-training reservation, and any applicable contractual terms collectively permit the intended

use of the content. Articles for which rights cannot be positively confirmed — whether due to missing metadata, ambiguous or conflicting licence designations, or unresolvable reservation signals — must be withheld from ingestion pending manual legal review. The default position in cases of uncertainty must be exclusion: the burden of demonstrating permissibility rests with the collection process, not with post hoc justification. This precautionary default is consistent with the principle of lawful processing under EU copyright law and mitigates the risk of inadvertent infringement at scale.

7.1.7 Lawful Access Channels and Statutory TDM Exceptions

Collection must occur exclusively through lawful access channels. Permissible mechanisms include institutional subscriptions, authorised APIs, and publicly accessible open repositories. Scraping content from behind authentication barriers, in contravention of website terms of service, or in violation of robots.txt policies is prohibited irrespective of whether the underlying content would otherwise qualify for inclusion based on its licence. The legality of the access mechanism is a condition independent of, and logically prior to, the licence-compatibility assessment; a permissively licensed article accessed through an unlawful channel does not acquire lawful status by virtue of its licence.

Where reliance is placed on the statutory TDM exceptions provided by Articles 3 and 4 of the CDSMD as the legal basis for access and processing, this reliance must be documented on a per-source basis. Documentation must include confirmation that access was initiated through a lawful channel, that the applicable exception (Article 3 for research organisations, or Article 4 for any person subject to the opt-out regime) is correctly invoked, and that no applicable opt-out was in force at the time of retrieval. This documentation forms part of the compliance record required under the EU AI Act and constitutes evidence of due diligence in the event of a third-party copyright claim.

7.2 Storage and access control

7.2.1 7Licence-Based Corpus Partitioning

The management of heterogeneous training corpora composed of scientific literature governed by different licensing regimes necessitates a storage architecture founded on strict partitioning. Training datasets must be physically and logically segregated into discrete storage domains that correspond to the licensing category of each constituent source. At minimum, the architecture shall distinguish three zones: an open-access zone, accommodating materials released under permissive licences such as CC0 or CC BY (including content sourced from OpenAlex or the PubMed Central Open Access subset), in which unrestricted processing and model training — including for commercial applications — is permitted subject to applicable attribution requirements; a non-commercial zone, housing content distributed under licences that prohibit commercial exploitation, such as CC BY-NC materials from PMC or DOAJ-indexed journals, governed by technical controls that prevent any workflow integration with commercially deployed AI systems; and a restricted or proprietary zone, reserved for content accessed under publisher text-and-data-mining (TDM) agreements or equivalent contractual instruments, in which use rights are defined exclusively by the terms of the underlying contract and must be documented at the dataset level.

Cross-domain contamination — the inadvertent processing of non-commercial or restricted content within pipelines governed by a more permissive regime — constitutes a licence-

compliance failure and must be structurally prevented. Each storage domain must therefore be assigned an explicit permissible-use policy that binds all downstream processing activities, including preprocessing, fine-tuning, and derivative dataset generation, to the rights profile of the originating corpus. This policy propagation requirement applies equally to intermediate artefacts produced during pipeline transformations, ensuring that licence constraints are preserved throughout the data lifecycle and are not diluted or lost as content moves between processing stages.

7.2.2 Permissible-Use Manifests and Policy Governance

Each storage zone must be associated with a formally documented or machine-readable use-policy manifest. This manifest shall specify, at minimum: the set of permitted processing operations applicable to the corpus; the authorised model-training use cases, differentiated where necessary between research, non-commercial, and commercial applications; any redistribution restrictions binding on derived outputs; and the applicable licence version or contractual reference. Maintaining such manifests in machine-readable form — for example, as structured metadata attached to dataset-level identifiers — facilitates programmatic enforcement during pipeline orchestration and reduces the risk of human error in policy interpretation.

The obligation to propagate use-policy rules to derivative artefacts is particularly significant in multi-stage preprocessing pipelines, where source content may be tokenised, filtered, deduplicated, or augmented before reaching the training environment. Each such transformation step must carry forward the most restrictive applicable licence constraint, applying a principle analogous to licence inheritance in software dependency management. Documentation of these constraints at every stage of the data pipeline is a precondition for demonstrating regulatory compliance in the event of an audit under the EU AI Act or an intellectual property dispute.

7.2.3 Access Controls for Restricted and Non-Commercial Corpora

Access to the non-commercial and restricted storage zones must be governed by role-based access control (RBAC) mechanisms that limit read and processing permissions to personnel whose operational functions are consistent with the terms of the applicable licence or TDM agreement. Access provisioning decisions must be documented and reviewed periodically to reflect changes in project scope, staffing, or contractual conditions. Where a TDM agreement identifies specific authorised users or research teams, access grants must correspond precisely to those designations.

All access events to restricted corpora must be logged with sufficient granularity to support retrospective auditing, licence verification, and regulatory inspections under the EU AI Act. Log records shall capture, at minimum, the identity of the accessing subject, the dataset or corpus partition accessed, the nature of the operation performed, and the timestamp of the event. These audit trails must be maintained for a retention period consistent with applicable contractual obligations and regulatory requirements, and must be protected against tampering or unauthorised deletion. The availability of comprehensive access logs is also a material element of accountability documentation for high-risk AI systems subject to the EU AI Act's conformity assessment requirements.

7.2.4 TRUST-Aligned Repository Governance

The selection of storage infrastructure for training corpora must be guided by the TRUST principles — Transparency, Responsibility, User Focus, Sustainability, and Technology — which constitute a recognised framework for the governance of digital repositories in research contexts. Applied to AI training data governance, these principles translate into concrete operational criteria: Transparency requires that storage governance structures, access conditions, dataset custodianship, and applicable use policies are documented and disclosed to all authorised stakeholders; Responsibility entails clear accountability for corpus integrity, licence compliance, and data-protection obligations; User Focus mandates that governance arrangements serve the legitimate needs of authorised researchers and engineers while constraining access in accordance with licensing terms; Sustainability requires long-term preservation commitments and continuity planning for training corpora, ensuring that provenance records and associated documentation remain accessible over the operational lifetime of the AI system; and Technology requires the deployment of infrastructure capable of enforcing access controls, maintaining audit trails, and supporting the technical safeguards mandated by applicable legal frameworks.

Repository selection must therefore prioritise infrastructures that provide explicit terms of use, stable and resolvable dataset identifiers, provenance metadata compliant with recognised standards, and demonstrated governance maturity consistent with TRUST certification or equivalent quality frameworks. Repositories lacking transparent governance structures or adequate provenance documentation are unsuitable for the storage of training corpora intended for use in regulated or commercially deployed AI systems.

7.2.5 Information Security Controls

Storage environments for training corpora must conform to the organisation's information-security standards across the full data lifecycle. Required controls include encryption of corpus data both at rest and in transit, integrity verification through cryptographic mechanisms such as hash-based validation applied at the dataset and file level, and systematic vulnerability management for all infrastructure components within the storage environment. These measures collectively ensure that corpus content is not susceptible to unauthorised access, alteration, or exfiltration, and that the integrity of training data — which directly affects the reliability and auditability of derived model behaviour — is verifiably maintained.

Security controls must be calibrated to the sensitivity classification of each corpus zone. The restricted zone and any zone containing personal data require heightened measures commensurate with the elevated risk profile of those datasets, including more stringent access authentication requirements, enhanced monitoring, and potentially physical or network-level isolation from general-purpose computing environments. The application of a tiered security model ensures that resource allocation for protective measures is proportionate to the actual risk presented by each category of corpus content.

7.2.6 Personal Data: Minimisation, De-identification, and Restricted Access

Scientific literature incorporated into training corpora may contain identifiable personal data, particularly in materials such as clinical case reports, qualitative research transcripts, patient-level observational studies, or other documents in which individuals are described with sufficient specificity to enable their identification. Where such content is present, the GDPR principle of data minimisation mandates that only the personal data strictly necessary for the intended AI training purpose is retained within the storage environment. This requirement implies that a

systematic review of corpus content must be conducted prior to storage ingestion, with the objective of identifying personal-data-bearing records and applying targeted exclusion or pre-storage de-identification rather than relying on post-hoc remediation, which is both operationally burdensome and less reliable as a compliance mechanism.

Where the complete exclusion of personal data from the corpus is not feasible, pseudonymisation or, where technically practicable, full de-identification must be applied prior to storage in training environments. Pseudonymisation reduces re-identification risk in a manner consistent with the safeguards required under GDPR Article 89 for the processing of personal data in scientific research contexts. Pseudonymisation procedures must be formally documented, including the key-management protocols that govern the retention and use of re-identification keys, which must be restricted to designated data-protection roles under strictly controlled conditions. The existence of re-identification keys within the same organisational perimeter as the pseudonymised corpus elevates the residual risk classification of the data and must be reflected in the applicable security measures and data-protection impact assessment.

Storage areas containing literature with residual personal data — whether pseudonymised or not yet fully de-identified — must operate under heightened access restrictions. Access must be limited to personnel who hold a documented legitimate research basis for processing the data and who have entered into formal data-access agreements specifying their obligations. In addition to RBAC-based access governance, technical controls including comprehensive audit trails, session-level monitoring, and automated alerts for anomalous access patterns must be implemented to enable timely detection and response to unauthorised access attempts. These controls serve both as preventive measures and as evidentiary resources in the event of a data-protection incident requiring notification to a supervisory authority under GDPR Article 33.

7.2.7 GDPR Research Safeguards and Dual Regulatory Compliance

The storage of personal-data-bearing scientific literature for AI training purposes must rest on a valid legal basis under the GDPR. Where scientific research exemptions are invoked pursuant to GDPR Article 89, the processing must be accompanied by the appropriate safeguards — specifically pseudonymisation, access restrictions, and data minimisation — and these safeguards must be demonstrably implemented and documented rather than merely asserted. The research exemption under Article 89 does not displace the requirement for a lawful basis under Article 6 (and, where special categories of data are involved, under Article 9), nor does it suspend the accountability obligations imposed on data controllers under Article 5(2). Storage governance documentation must accordingly reflect both the legal basis relied upon and the specific safeguards applied to personal data within the training environment.

For AI systems classified as high-risk under the EU AI Act, GDPR compliance obligations co-exist with the data-governance requirements of Articles 10 and 17 of that Regulation, which impose obligations relating to training data quality, relevance, and the management of known biases. Storage architectures must be designed to satisfy both regulatory frameworks simultaneously, without generating conflicts or gaps in compliance coverage. In practice, this requires that data governance policies addressing GDPR obligations — including records of processing activities, data-protection impact assessments, and documentation of applicable safeguards — are integrated into the broader technical documentation required for EU AI Act conformity, rather than maintained as separate and potentially inconsistent instruments. The convergence of these two regulatory obligations reinforces the need for a unified, systematically maintained

governance framework for the storage and access control of AI training data derived from scientific literature.

7.3 Processing and preparation for LLM use

7.3.1 Processing Logs and Traceability

The governance of training corpora for large language models requires that every transformation applied to the dataset be systematically recorded, from initial ingestion through all intermediate processing stages to the final corpus state used in a training run. End-to-end logging obligations thus mandate the capture of each operation performed — including ingestion filters, deduplication procedures, normalisation steps, and format conversions — so that the complete lineage of any derived dataset may be reconstructed upon request by regulators, auditors, or rightsholders.

At the stage level, log entries must identify not only the operations performed but also the specific sources included in each processing batch and the licence terms applicable to those sources at the time of processing. This dual requirement ensures post-hoc verifiability of compliance decisions and precludes any ambiguity arising from subsequent changes to licence conditions or source availability. Licence metadata must travel with the data through all transformation stages; no downstream representation — whether a filtered subset, a tokenised corpus, or an embedded representation — may be legally or technically dissociated from the source-level rights information that governs its use.

Processed corpus versions must be identified through persistent and reproducible fingerprints. Cryptographic hashes — such as SHA-256 digests applied to deterministically serialised corpus files — or structured dataset identifiers provide a mechanism for verifying that a specific, immutable dataset state was used in a given model training run. Where scientific literature constitutes part of the corpus, persistent scholarly identifiers such as Digital Object Identifiers (DOIs), ORCID contributor identifiers, and Research Organisation Registry (ROR) identifiers should be propagated into corpus metadata to preserve provenance at the granularity of individual documents and contributors.

The processing infrastructure must additionally be designed to support targeted exclusion or removal of specific data items without requiring full corpus reconstruction. This capability is operationally necessary to respond to rightsholder takedown requests, regulatory instructions, or the post-hoc identification of impermissible content. Comprehensive audit trails linking log entries to identifiable data items serve a further function under the EU AI Act conformity assessment process: they enable the organisation to demonstrate compliance to enforcement authorities and notified bodies through traceable documentary evidence rather than retrospective reconstruction.

7.3.2 Data Quality Assessment

Data quality governance for AI training corpora should be implemented in alignment with the ISO/IEC 5259 series, which establishes process models and standardised metrics for evaluating quality across the full machine-learning data lifecycle. Compliance with this framework requires quality management processes that address both static dataset properties — including completeness, accuracy, and internal consistency — and dynamic concerns such as version drift

and progressive contamination from synthetic or low-quality content as the corpus evolves over time.

Assessment of domain completeness requires that the corpus provide adequate coverage of the intended application field, whilst statistical representativeness demands coverage across the relevant population subgroups as specified under Article 10 of the EU AI Act. In the context of medical scientific literature, representativeness considerations extend to clinical subspecialties, patient demographics, geographic origin of studies, and publication language, each of which may affect downstream model behaviour in clinically significant ways.

Error-rate evaluation constitutes a further dimension of quality assessment. This encompasses the systematic identification and quantification of retracted articles, expressions of concern, confirmed duplicate records, formatting artefacts introduced by optical character recognition or conversion pipelines, and machine-generated or otherwise low-quality content. Each of these categories introduces distinct degradation pathways into the trained model and must therefore be addressed through documented remediation processes. In accordance with the continuous improvement requirements of the ISO/IEC 5259 framework, each identified quality defect must trigger a documented remediation action — such as record exclusion, metadata correction, or controlled re-ingestion — with the outcome recorded alongside a reference to the original quality assessment finding.

7.3.3 Bias Assessment and Mitigation

Providers of high-risk AI systems are subject to an explicit obligation under Article 10 of the EU AI Act to conduct systematic examinations of training datasets for potential biases. In the domain of medical literature, bias may arise from skewed coverage of clinical domains, geographic or linguistic concentration of published research, structural underrepresentation of specific demographic groups in clinical studies, or funding-related publication patterns that favour particular therapeutic approaches. Bias assessment must be applied not only to the raw corpus at the point of initial ingestion, but also to any filtered or processed subsets derived from it, since selection operations and pre-processing decisions may independently introduce or amplify representational imbalances that were not present in the source data.

All bias mitigation strategies applied to the corpus must be documented with sufficient technical specificity to permit replication and to withstand regulatory scrutiny. Documented measures may include oversampling of underrepresented source categories, stratified sampling by language, geography, or clinical domain, language-balanced filtering to ensure multilingual coverage, or the explicit exclusion of source collections identified as systematically unrepresentative. Where residual biases cannot be fully remediated — whether due to structural limitations of the available literature or to resource constraints — such residual biases must be disclosed in the dataset documentation, in alignment with the transparency requirements of established frameworks including Datasheets for Datasets, Data Cards, and the STANDING Together consensus recommendations applicable to health-domain datasets.

7.3.4 Use-Case Boundary Enforcement and Licence Propagation

A fundamental requirement of lawful corpus management is the strict enforcement of use-case boundaries between data partitions collected under different authorisation regimes. Content acquired under research-only or non-commercial text and data mining (TDM) licences must be physically or logically segregated from corpora intended for commercial model training, and governance policies must designate permissible uses for each partition with corresponding

technical access controls. Such controls must prevent cross-partition use — whether intentional or incidental — that would violate source-level licence restrictions or statutory TDM opt-outs exercised by rightsholders under Article 4(3) of Directive (EU) 2019/790.

In concrete terms, content ingested under Creative Commons licences carrying a non-commercial condition — including CC BY-NC, CC BY-NC-SA, and CC BY-NC-ND — or under non-commercial TDM agreements such as those governing access to the Elsevier TDM API, must be excluded from any training pipeline whose output is intended for commercial deployment or for distribution as a commercial AI product. Internal governance controls must further prevent pipeline reconfiguration or corpus reuse that would expose commercially restricted content to commercial training workloads, including through indirect use in intermediate representation learning tasks that subsequently contribute to commercially deployed model weights.

Beyond use-case boundary enforcement, licence conditions attaching to training data must be propagated into the terms governing downstream model distribution, API access, and derivative product licensing. Conditions of particular significance include share-alike obligations arising under CC BY-SA and CC BY-NC-SA licences, non-derivative restrictions, and AI-training reservations that may be embedded in individual licence agreements or exercised through machine-readable opt-out mechanisms. For general-purpose AI models covered by the EU AI Act, this propagation obligation is complemented by the requirement to publish a sufficiently detailed summary of copyright-protected training data categories, enabling rightsholders to verify that their opt-out declarations have been respected and to exercise any remedies available under applicable copyright law.

7.4 Sharing, retention, and deletion

The governance of training corpora does not terminate upon model deployment. It extends, with equal legal force, to the subsequent sharing of training data, the conditions under which retained materials may be stored, and the mechanisms by which specific sources may be removed or excluded. These obligations arise from the intersection of intellectual property law, contractual licence terms, the EU AI Act's requirements for general-purpose AI (GPAI) model providers, and emergent best practices in data documentation. This section addresses each dimension in turn, with particular attention to the medical scientific literature context that characterises the present project.

7.4.1 Permissibility of External Data Release

The permissibility of releasing any component of a training corpus to third parties is strictly conditioned by the licence terms applicable to each constituent source. No external release is lawful without prior verification that the governing licence expressly authorises redistribution, and the analysis must be conducted at source granularity rather than at corpus level, given that heterogeneous corpora routinely integrate materials under materially different licensing regimes.

Open-licensed content released under Creative Commons Zero (CC0) or Creative Commons Attribution (CC BY) licences may be redistributed, including in commercial contexts, subject to compliance with applicable attribution obligations. Where share-alike variants apply — notably CC BY-SA — any derivative release is additionally constrained by reciprocal licensing requirements, which may propagate to the downstream distribution terms of models trained on

such content. Content licensed under CC BY-NC or CC BY-NC-ND presents a more restrictive profile: it must be withheld from any commercial release and may only be shared within the scope of non-commercial or research-limited deployments, conditions that necessitate careful classification of the intended use prior to any disclosure.

Proprietary or rights-reserved content — encompassing subscription publisher material, paywalled corpora, and sources carrying explicit machine-learning or AI-training reservations — must be excluded from any external distribution absent specific contractual authorisation. In the context of medical scientific literature, this category is particularly significant, as a substantial proportion of journal content remains subject to publisher licences that either prohibit redistribution entirely or impose conditions inconsistent with general corpus release. The presence of machine-readable opt-out signals, including those communicated through Crossref Text and Data Mining (TDM) services or encoded in robots.txt directives, constitutes an additional restriction that is legally operative under EU TDM provisions and must be respected irrespective of the open-access status of the work in question.

7.4.2 Sharing Protocols for Mixed-Licence Corpora

Where a training corpus integrates content under heterogeneous licence regimes — a configuration that is typical of large-scale scientific literature collections — external sharing must be disaggregated according to the licensing profile of each constituent subset. Full-text release is only permissible for subsets governed by licences that expressly authorise redistribution; CC0 and CC BY represent the primary categories satisfying this condition. Subsets subject to CC BY-SA may be released subject to compliance with the share-alike obligation, while subsets under NC or ND variants must be withheld from any commercial release pipeline.

For corpora in which unrestricted release of full-text subsets is not feasible across the entire collection, the appropriate instrument is the release of non-displacive metadata, bibliographic references, or aggregated statistical indicators that do not reproduce or enable reconstruction of protected content. Aggregate outputs — such as frequency distributions, domain coverage metrics, vocabulary statistics, or corpus composition profiles — may be disclosed without implicating reproduction rights, provided that the disclosed information does not function as a functional substitute for the underlying works. This approach enables meaningful transparency regarding corpus characteristics while maintaining compliance with rights-reserved source terms.

7.4.3 GPAI Transparency and Training Data Summaries

Providers of general-purpose AI models within the scope of the EU AI Act are subject to a specific obligation to prepare and publish sufficiently detailed summaries of the copyright-protected training data categories and major sources used in model development. This obligation, which arises under Article 53 of the AI Act and the associated GPAI model transparency requirements, is designed to enable rightsholders to verify whether their works were incorporated into a training corpus and whether applicable TDM opt-outs were respected.

The publication obligation does not require full corpus disclosure and is not equivalent to an obligation to release the training data itself. It is satisfied by structured, source-level categorisation accompanied by attestations of licence compliance and opt-out adherence. The summary must be sufficiently granular to enable meaningful verification by rightsholders, which

in practice implies disclosure at the level of major source categories — for instance, distinguishing between open-access repository content, licensed database content, and web-harvested material — together with an indication of the licence or rights framework governing each category. Generic or uninformative summaries that do not enable rightsholder verification would not satisfy the statutory standard.

7.4.4 Corpus Documentation Standards

Beyond regulatory disclosure obligations, the documentation of training corpora has become a recognised discipline with standardised artefact formats, each serving distinct but complementary purposes. Datasheets for Datasets constitute the baseline transparency instrument: they capture the motivation for dataset creation, its composition, the collection methodology, pre-processing and transformation operations applied, intended and foreseeable uses, and known legal or ethical constraints. In the context of medical literature corpora, the clinical and biomedical domain introduces additional considerations regarding data sensitivity, potential bias in source population representation, and limitations on permissible downstream applications.

Data Cards and Dataset Nutrition Labels provide structured, human-readable summaries presenting provenance, known limitations, and permissible uses, offering richer contextual metadata than the baseline Datasheet format and facilitating rapid assessment by downstream users or compliance reviewers. Data Portraits represent a more technically sophisticated instrument: they are machine-readable records enabling downstream users to query which specific training examples were included in a model corpus, thereby supporting rightsholder verification requests and enabling targeted content removal without requiring full corpus reconstruction. Open Datasheets, a related format, encode responsible AI considerations in machine-readable form and are designed to support automated risk and compliance assessment pipelines.

Documentation released alongside any training corpus must explicitly identify which portions derive from openly licensed infrastructures — such as OpenAlex, the PubMed Central Open Access Subset, or Wikimedia — and which derive from restricted, licensed, or proprietary sources. The recommended instrument for this disclosure is a licensing and rights matrix specifying, for each source, the applicable licence variant, TDM and AI-training permissibility, any active opt-outs, and the jurisdictional assumptions underlying the compliance analysis. Physical and logical segregation of open and restricted subsets within storage and documentation structures reinforces both compliance posture and audit readiness.

7.4.5 Lifecycle Traceability and Audit Infrastructure

Comprehensive lifecycle traceability requires maintaining versioned processing logs, dataset snapshots with timestamps, and persistent identifier linkage — through DOIs, ORCID researcher identifiers, and ROR institutional identifiers — at each stage of collection, transformation, and model integration. This infrastructure serves multiple legal functions: it enables demonstration of licence compliance at the time of corpus construction, supports audit in the event of a regulatory inquiry or rightsholder challenge, and provides the factual basis for targeted content removal in response to post-hoc opt-out assertions or legal claims.

Hashes or content identifiers for derived datasets must be recorded alongside their source lineage to support downstream verification and selective exclusion. In operational terms, this requires that each ingestion, transformation, or filtering step be logged in a manner that preserves the relationship between derived artefacts and their original sources, such that any specific document or source can be identified and excluded without disrupting the integrity of the broader dataset. This capability is not merely a best-practice recommendation; in the medical domain, where individual publications may carry privacy-sensitive information or be subject to updated publisher policies, the ability to execute precise removal operations is a functional requirement of responsible corpus management.

7.4.6 Takedown, Exclusion, and Source Blocking Mechanisms

Governance frameworks for AI training data must operationalise the capacity to identify, isolate, and remove contributions from specific sources — at both the dataset level and, where technically feasible, at the model-training level — in response to legitimate legal claims, regulatory directives, or updated rightsholder instructions. This operationalisation requires source-level logging combined with granular provenance metadata, enabling targeted responses to individual takedown or exclusion requests without requiring reconstruction of the entire corpus.

Blocking mechanisms for specific sources must be integrated into the ingestion pipeline to prevent the reintroduction of excluded content in subsequent training iterations. This is a non-trivial architectural requirement: without explicit pipeline controls, excluded material may be inadvertently reingested when corpora are refreshed or augmented, potentially reinstating a compliance violation that had previously been remediated. In practice, this implies maintaining a persistent exclusion registry — keyed on source identifiers, DOIs, or publisher domains — that is consulted at each ingestion event and enforced as a mandatory pre-processing gate.

The technical architecture of the exclusion and takedown system must also account for the distinction between dataset-level and model-level removal. Dataset-level removal — the deletion or exclusion of specific content from the training corpus — is straightforward to implement and constitutes the primary compliance mechanism. Model-level removal, by contrast, raises unresolved technical questions regarding machine unlearning and the extent to which trained model weights can be modified to eliminate the influence of specific training examples; this remains an active area of research and is not currently a settled legal obligation, though it is increasingly discussed in regulatory contexts.

7.4.7 Ongoing Policy Review and Governance Maintenance

Data sourcing and licensing policies must be treated as living governance instruments subject to periodic reassessment rather than static compliance determinations made at the point of corpus construction. Publisher reservation practices evolve — as evidenced by the progressive introduction of AI-training opt-out clauses across major scientific publishers — and new open corpora periodically emerge that may alter the optimal composition strategy for a given project. Regulatory and standards guidance likewise continues to develop, with implementing acts under the EU AI Act, guidance from national data protection authorities, and emerging technical standards all potentially affecting the permissibility of previously accepted practices.

Review cycles should incorporate systematic monitoring of machine-readable opt-out signals, including those communicated through Crossref TDM services and publisher-level policy announcements, ensuring that sources whose terms have changed are re-evaluated against current permissibility standards. Where a previously accepted source is found to have introduced new restrictions — for example, by adding an AI-training exclusion to its terms of use — the governance framework must include a defined response protocol specifying the timeline and procedure for exclusion from active and future training runs.

All policy updates must be documented and versioned to maintain an auditable record of governance decisions over the corpus lifecycle. This requirement serves both internal accountability purposes and external compliance demonstration: in the event of a regulatory inquiry or legal challenge, a versioned policy record enables the data controller to demonstrate that governance decisions were made in good faith, based on the information available at the relevant time, and in accordance with established review procedures. The combination of lifecycle traceability infrastructure, documented exclusion mechanisms, and versioned policy records constitutes the minimum governance architecture necessary for defensible and sustainable compliance in the context of AI training on medical scientific literature.

8 Governance, roles, and responsibilities

8.1 Roles within the consortium

The governance of any consortium engaged in AI development using scientific literature as training data requires the establishment of a clearly defined and hierarchically structured role framework. Given the intersection of intellectual property law, data protection regulation, research ethics, and technical engineering that characterizes such projects, no single organizational unit can bear undivided responsibility. Instead, accountability, decision-making authority, and operational obligations must be distributed across specialized roles at the strategic, tactical, and operational levels. The present section sets out the allocation of roles within the consortium, describing the responsibilities and interrelationships of each function in a manner consistent with the requirements of applicable governance frameworks, including NIST AI RMF and ISO/IEC 42001.

Table 7. Governance matrix for data-source compliance decisions. R = Responsible; A = Accountable; C = Consulted; I = Informed.

Activity / decision	Source assessor	Legal / IP lead	Technical lead	WP2 lead	Consortium management	Record / evidence required
Identification of candidate sources	R	I	I	I	I	Initial source reference and provenance note
Preliminary legal eligibility screening	C	R	I	A	I	Rights and eligibility assessment record

Review of licence terms and contractual restrictions	I	R	I	A	I	Licence review note or contractual interpretation record
Assessment of personal or health-related data implications	C	R	C	A	I	GDPR and safeguards assessment note
Technical assessment of feasibility for ingestion and segregation	I	C	R	A	I	Technical handling and segregation note
Decision on acceptance, rejection, or conditional use	C	R	C	A	I	Formal decision record with rationale
Escalation of ambiguous or high-risk cases	I	R	C	A	C	Escalation note and decision trail
Validation of documentation and traceability completeness	C	R	C	A	I	Compliance file and audit trail entry
Periodic review of previously accepted sources	C	R	C	A	I	Review update and status confirmation
Final oversight for exceptional or consortium-wide risk cases	I	C	I	R	A	Management decision note where required

8.1.1 Strategic and Executive Layer

Executive Sponsor and Board-Level Accountability

Ultimate accountability for AI risk exposure resides at board level. The Executive Sponsor occupies the apex of the governance hierarchy, bearing institutional responsibility for ensuring that AI initiatives remain aligned with the organisation's strategic objectives and that the organisation's risk exposure — including legal, reputational, and regulatory risk — is managed within defined tolerances. This encompasses oversight of the full AI lifecycle: from the lawfulness of training data acquisition and compliance with applicable intellectual property frameworks through to potential regulatory sanctions arising from non-compliant use of copyright-protected or personal data.

The Executive Sponsor does not engage in day-to-day operational governance, but exercises authority as the final escalation point for cross-functional conflicts that cannot be resolved at

lower layers of the hierarchy. This role is therefore structural rather than operational, providing the institutional weight necessary to enforce decisions that span organisational boundaries or require trade-offs between competing institutional mandates.

AI Governance Committee

Collective oversight over AI risk policy is exercised by a dedicated AI Governance Committee, which functions as the principal decision-making body for high-risk or legally complex AI projects. In the context of projects involving the ingestion of copyright-protected scientific literature, the processing of personal data derived from clinical or genomic sources, or the use of dual-use content, the committee is responsible for setting the organisation's risk appetite and defining the conditions under which such material may lawfully be incorporated into training corpora.

The committee constitutes the formal escalation pathway for unresolved compliance, ethical, or reputational risks arising at lower governance layers. It also holds authority over the approval or suspension of active literature ingestion pipelines where changes in law, publisher policy, or institutional risk classification so require. To fulfil its mandate effectively, the committee must operate on the basis of informed cross-functional deliberation, drawing on specialist expertise from all domains with material exposure to AI risk.

Cross-Functional Composition of the Strategic Layer

Effective governance at the strategic layer is contingent on the breadth and depth of cross-functional representation. The AI Governance Committee must include intellectual property and copyright counsel, data protection officers, information security specialists, research ethics representatives, library or licensing professionals with subject-matter expertise in publisher contracts and Text and Data Mining (TDM) frameworks, and product ownership representatives who can articulate project-level risk tolerance. This composition ensures that policy decisions on training data reflect the full spectrum of regulatory obligations — encompassing TDM compliance, privacy law, contractual terms with scientific publishers, and downstream product risk — rather than the perspective of any single function. In consortium settings where multiple institutional mandates intersect, such diversity of membership is essential for holistic risk assessment and the avoidance of siloed decision-making.

8.1.2 Tactical and Programme Layer

AI Governance Manager and Programme Owner

The AI Governance Manager is responsible for operationalising the governance framework established at the strategic layer. This role serves as the primary interface between strategic policy and project-level execution, translating high-level governance requirements into actionable workflows, documentation obligations, and control mechanisms applicable to active AI development projects. Specific responsibilities include coordinating risk assessments, maintaining the inventory of AI systems subject to governance oversight, and ensuring that literature-based training projects adhere to the controls approved by the AI Governance Committee.

The AI Governance Manager also bears responsibility for horizon-scanning: monitoring for material changes in applicable law, publisher policy, or institutional risk classification that may necessitate review or suspension of active ingestion pipelines. This function is ongoing rather than episodic, given the dynamic regulatory environment surrounding AI, data protection, and intellectual property at both the national and EU level.

Data Governance Office

The Data Governance Office owns the organisation's data governance framework, including the data catalogue, lineage infrastructure, and versioning controls. In the context of AI training using scientific literature, the remit of the Data Governance Office extends to cover all AI training datasets derived from corpus ingestion. It is responsible for defining and enforcing standards for dataset documentation, data quality assessment, and provenance tracking, and constitutes the institutional memory for all corpus-level governance decisions.

Critically, the Data Governance Office provides the shared tooling and procedural infrastructure required to ensure reproducibility, post-hoc auditability, and regulatory transparency for literature-derived corpora across multiple AI projects. This capacity is of particular importance where regulatory inquiry or litigation may require the reconstruction of the dataset composition at a specific point in time, or where changes to a dataset must be traced to a specific governance decision.

Ethical AI and Responsible AI Function

A dedicated Ethical AI or Responsible AI function is responsible for designing and maintaining the fairness, accountability, and transparency practices applicable to training data workflows. This includes the development and stewardship of documentation standards such as data cards and model cards, which serve as instruments of both internal governance accountability and external regulatory transparency. The function conducts structured review of high-risk AI systems for bias exposure, dual-use implications, and alignment with disciplinary norms — an obligation of particular relevance where scientific literature may encode domain-specific representation imbalances that, if unaddressed, could compromise the reliability or scientific integrity of the resulting model.

The Ethical AI function operates as an institutional counterweight to purely technical or commercial considerations, ensuring that governance decisions reflect broader ethical commitments and the expectations of the scientific and clinical communities whose published work forms the basis of the training corpus.

8.1.3 Operational and Project Layer

Project Lead and Product Owner

At the project level, the Project Lead or Product Owner bears direct accountability for the AI system's performance objectives and for compliance with the governance controls prescribed for the specific project. This role constitutes the primary point of operational responsibility, ensuring that model development, corpus selection, and deployment decisions remain within the boundaries established by the governance and risk framework approved at higher layers. The Project Lead is responsible for ensuring that project-level decisions are escalated appropriately where they approach or exceed established risk thresholds.

Corpus Data Owner

A named Corpus Data Owner must be designated for each major literature-derived dataset used in AI training. This individual is responsible for the full lifecycle of the dataset: source selection and acquisition, verification and documentation of lawful access, maintenance of licensing records, and authorisation of any modification to the dataset's composition. The Corpus Data Owner constitutes the principal point of accountability for intellectual property and access

compliance at the dataset level, providing an auditable chain of custodianship from initial acquisition through to deployment in model training.

The role carries ongoing obligations. Changes in publisher terms, rights-holder opt-outs, or emerging legal interpretations of TDM exception eligibility must be promptly reflected in the dataset's status record and acted upon — including, where necessary, the exclusion of affected materials from active training pipelines. The Corpus Data Owner therefore operates at the intersection of legal compliance, data governance, and operational data management.

Data Stewards and Engineers

Data Stewards and Data Engineers are responsible for implementing the technical infrastructure through which governance policy is translated into operational reality. Their responsibilities encompass pipeline design and execution, the application of de-identification or anonymisation procedures where required by data protection obligations, and the recording of data lineage in a manner that supports both internal auditability and external regulatory disclosure. Data Stewards are specifically responsible for populating dataset documentation artefacts — including data cards — and for maintaining the accuracy of catalogue entries throughout the dataset lifecycle.

Domain Experts

Domain experts — including scientists and clinicians with subject-matter expertise relevant to the corpus in question — provide the domain-specific quality assurance that cannot be substituted by technical personnel. Their role within the governance structure is to validate the scientific suitability of curation rules, annotation schemes, and exclusion criteria applied to the literature corpus, and to review the bias profile and representational coverage of the corpus relative to the intended AI application. Domain experts are positioned to identify disciplinary blind spots, imbalances in the representation of research traditions or populations, or deficiencies in source selection that would be invisible to technical reviewers without specialist knowledge. Their input is therefore a necessary condition for responsible corpus governance rather than an optional enhancement.

Legal, Intellectual Property, and Data Protection Officers

Legal and Intellectual Property Officers and Data Protection Officers provide binding interpretations of copyright law, database rights, publisher contracts, and privacy regulation as applied to specific datasets and ingestion decisions. They constitute the authoritative legal voice within the project governance structure, and their sign-off is required for any ingestion decision involving significant legal uncertainty — including, in particular, cases where TDM exception eligibility is disputed, where publisher contractual terms are ambiguous, or where personal data may be present in clinical, genomic, or patient-record content.

These officers serve as the formal interface between project-level operations and the regulatory and contractual obligations that govern the lawful use of scientific literature for AI training. They do not perform operational functions but are consulted and, for high-risk decisions, hold co-accountability with the Corpus Data Owner and the AI Governance Committee.

8.1.4 RACI Allocation of Responsibilities

Governance frameworks aligned with NIST AI RMF and ISO/IEC 42001 require explicit assignment of accountability, responsibility, consultation, and information (RACI) obligations for all key

activities in AI projects. This ensures that decision authority, operational execution, specialist input, and information flows are unambiguous, consistently applied, and auditable across the full project lifecycle. For literature-based AI training projects, RACI allocation must cover at minimum the following activity domains: (i) approval of scientific literature use for AI training; (ii) definition of source selection criteria and acquisition methods; (iii) assessment of copyright, TDM exception eligibility, and licensing compliance; (iv) privacy and ethics risk assessment; (v) corpus ingestion approval; (vi) data catalogue and lineage maintenance; and (vii) model deployment authorisation.

The distribution of Accountable (A), Responsible (R), Consulted (C), and Informed (I) designations across the Executive Sponsor, AI Governance Committee, Data Governance Office, Project Lead, Corpus Data Owner, Legal/IP Officers, Data Protection Officers, and Domain Experts must reflect both the governance hierarchy established above and the technical and legal specialisation of each role. In particular, accountability must not be over-concentrated in a single role — whether at the operational level, where this would create bottlenecks and single points of failure, or at the executive level, where it would divorce accountability from the operational knowledge required to exercise it meaningfully.

The explicit RACI assignment across these roles operationalises the 'Govern' function of the NIST AI RMF and satisfies the roles-and-responsibilities requirements of ISO/IEC 42001:2023. It provides the institutional foundation necessary to ensure that AI governance decisions are consistent, traceable, and legally defensible — conditions that are especially demanding in the context of training data derived from copyright-protected scientific literature, where both the regulatory landscape and the contractual environment continue to evolve.

8.2 Decision-making and escalation process

The governance of AI systems that ingest scientific literature for training purposes requires a formalised decision-making architecture capable of addressing legal, ethical, and operational risks in a structured and auditable manner. This section defines the procedural framework through which data-intake decisions are assessed, authorised, escalated, and monitored over time. The process is not conceived as a single point-in-time gate but as a continuous governance cycle, consistent with the requirements of the NIST AI Risk Management Framework (AI RMF) Govern function, ISO/IEC 42001 leadership and responsibility controls, and applicable EU regulatory instruments, including the EU AI Act and the Copyright in the Digital Single Market (CDSM) Directive.

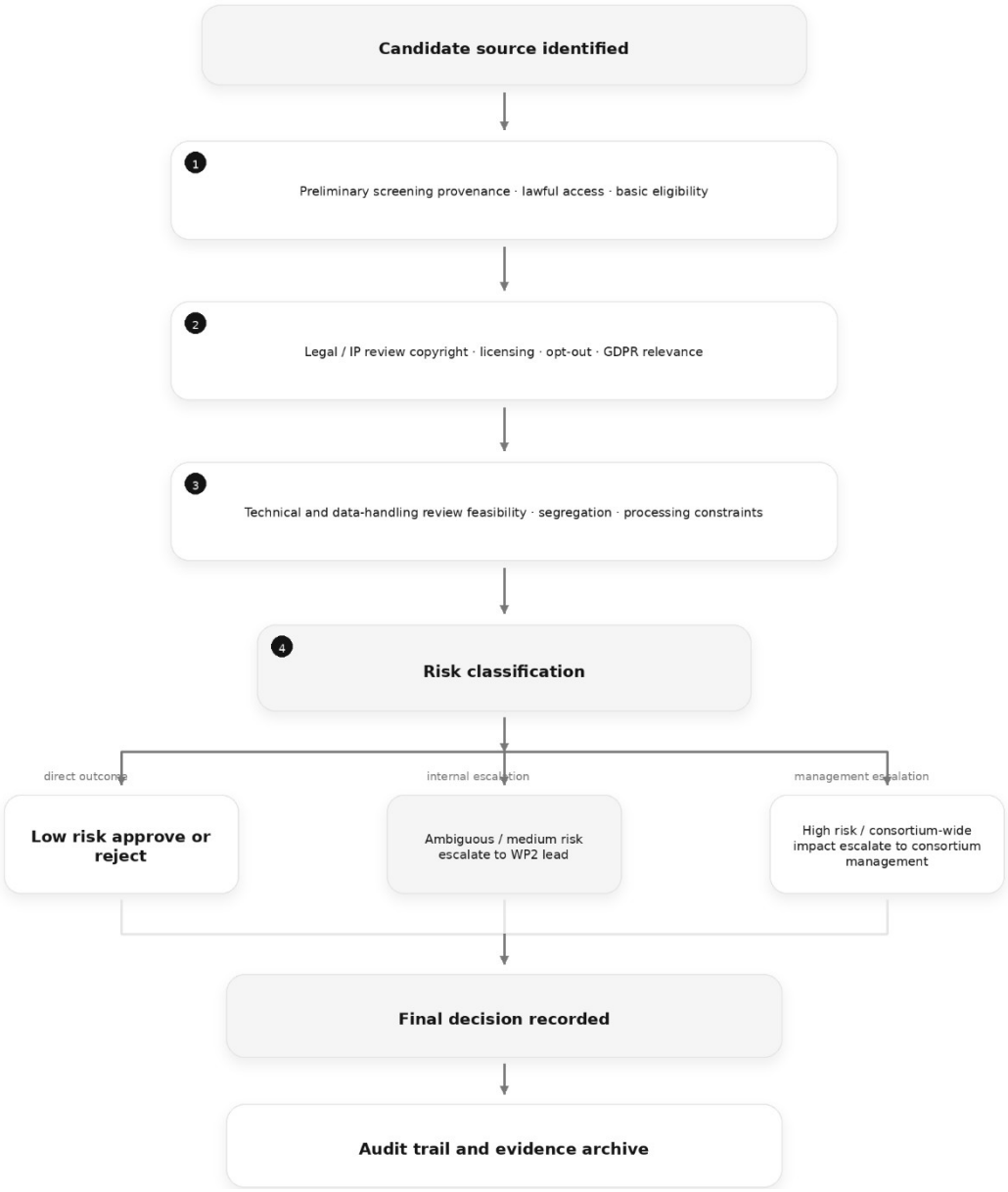

Figure 4. Risk register for the use of medical literature in LLM training

8.2.1 Governance structure and accountability

Effective governance of AI data intake cannot be delegated to individual project teams operating in isolation. The decision-making architecture must therefore be operationalised through a multi-layer accountability structure that encompasses executive sponsorship at the organisational level, cross-functional committee oversight at the programme level, and clearly designated project-level ownership for day-to-day compliance obligations. Each layer must hold defined authority over risk thresholds and data-intake decisions proportionate to the sensitivity and complexity of the processing involved.

Governance policies must explicitly codify who bears responsibility for AI risk, how risk trade-offs are adjudicated between legal, technical, and scientific objectives, and how AI governance processes integrate with the organisation's enterprise-wide risk management framework.

Decision processes for literature-based AI projects must be formalised as standing procedures — not ad hoc judgements — covering ingestion approval, legal basis confirmation, escalation pathways, and documentation obligations. This formalisation ensures that governance is repeatable, auditable, and resistant to dependency on particular individuals.

8.2.2 Pre-ingestion legal and rights assessment

Prior to the ingestion of any new literature corpus, a structured legal basis assessment must be conducted and documented as a formal precondition to data-intake approval. This assessment must identify the applicable legal ground — whether open-access licence, institutional subscription, contractual agreement, or a statutory text and data mining (TDM) exception — for each intended use of the corpus, including training, fine-tuning, evaluation, and any prospective commercialisation. Legal basis analysis must be linked to the dataset record in the organisation's data catalogue and must explicitly flag any use for which a valid legal foundation cannot be established.

The copyright and database-right review must determine whether the planned reproduction and extraction activities are covered by lawful access, a valid licence, or a statutory TDM exception under the applicable jurisdiction. The distinction between commercial and non-commercial use scenarios is material: reliance on a non-commercial TDM exception under Article 3 of the CDSM Directive will not extend to subsequent commercialisation of a model trained on that corpus. The lawful-acquisition standard — ensuring that source material has been legitimately obtained before ingestion commences — constitutes the primary governance lever for copyright risk management. US litigation precedent confirms that training on infringing material can generate statutory copyright damages even where downstream fair use arguments might otherwise succeed, underscoring the importance of upstream verification.

In parallel with copyright analysis, publisher contracts, platform terms of service, and dataset licence conditions must be reviewed to identify restrictions on automated harvesting, bulk export, or AI training use. Contractual constraints must be recorded in the data catalogue alongside the copyright assessment and must be flagged for re-assessment whenever the AI system's deployment scope or commercialisation status changes. Contractual restrictions may be more stringent than applicable statutory law and must be treated as binding obligations independent of any fair use or TDM exception analysis.

8.2.3 Data protection, privacy, and sectoral ethics review

Where a literature corpus contains personal or sensitive data — including patient narratives, genomic identifiers, or author-linked clinical case reports — a data-protection review must be conducted to identify the applicable legal basis under the General Data Protection Regulation (GDPR) or equivalent legislation and to assess the proportionality of the intended processing. Where personal data is present, mandatory safeguards include de-identification or redaction pipelines, data protection impact assessments (DPIAs) for high-risk processing activities, and access-control restrictions commensurate with data sensitivity. The Data Protection Officer (DPO) holds accountable authority over privacy risk determination; this responsibility may not be delegated to the project team alone, and the DPO must be re-consulted whenever the processing purpose or system deployment context changes materially.

In regulated domains such as biomedical research, genomics, and digital pathology, domain-specific ethical and regulatory frameworks — including human-subjects protection rules and biomedical research ethics standards — may impose constraints on data use that operate

independently of copyright permissions. Sectoral ethics review must therefore be treated as a mandatory gate in the data-intake process for any corpus derived from clinical, experimental, or human-subjects research contexts. This review must be conducted by a dedicated responsible-AI or institutional review function, and its outcome must be recorded as part of the formal ingestion approval decision.

A dedicated risk and ethics review must further assess corpus-level risks including geographic, linguistic, or disciplinary overrepresentation that could amplify existing inequities in research attention; dual-use risks arising from sensitive methodological content in fields such as virology or synthetic biology; and alignment with community attribution and scholarly-credit norms. Scientific-integrity risks — including the potential for models trained on uncurated corpora to generate hallucinated citations or fabricate evidence claims — must be identified at the data-intake stage and addressed through curation constraints, filtering controls, and human-in-the-loop validation requirements.

8.2.4 Use-case classification, source mapping, and TDM opt-out verification

The intended AI system and its use case must be classified according to applicable regulatory and institutional risk frameworks — including the EU AI Act's high-risk categories and the designation of general-purpose AI (GPAI) models — before data sourcing commences. Risk classification determines the intensity of governance required at each stage: a change in deployment scope from an internal research tool to a commercial product, or from general use to clinical decision-support, constitutes a material change that triggers re-classification and a full reassessment of the data governance basis.

All literature sources must be enumerated prior to ingestion, with each source characterised by its access mode (open access, subscription API, web crawl, or institutional repository dump), the applicable licence or statutory exception, and any detected rights-holder opt-out status for commercial TDM. This source mapping must be maintained as a living registry within the data catalogue, updated whenever new sources are added, existing licence terms change, or opt-out notices are detected.

Under the CDSM Directive, TDM of lawfully accessed works is permitted, but rights holders may reserve their rights for commercial TDM by means of machine-readable opt-out notices. The EU AI Act extends this obligation to GPAI model providers, requiring both documented compliance with TDM rules and the publication of a sufficiently detailed summary of copyrighted training content. Verification of opt-out status must accordingly be integrated into the source-mapping workflow, with automated or periodic checks for newly issued opt-out notices that may affect previously approved corpora.

8.2.5 Formal legal, IP, and DPO review and sign-off

Legal and intellectual property (IP) counsel must formally assess the copyright, database rights, and contractual position for each planned use of a corpus — including training, evaluation, fine-tuning, and commercialisation — and must provide sign-off on high-risk ingestion decisions as a binding condition of approval. Legal review must identify where specific licences, alternative sources, or contractual amendments are required to support the intended use, and must flag any uses for which no adequate legal basis exists. This review may not be substituted by project-team self-assessment for decisions involving novel legal bases, sensitive domains, or significant legal uncertainty.

The DPO must independently evaluate each corpus for the presence of personal data and determine whether processing is permissible under applicable data-protection law, taking into account the principles of purpose limitation, data minimisation, and lawful basis. The DPO must specify required technical and organisational safeguards as binding conditions on ingestion approval.

8.2.6 Decision record and conditional approval

Every ingestion decision must be recorded as one of three formal outcomes: approve, approve with conditions, or reject. The decision record must document the rationale with reference to the applicable legal, ethical, and technical assessments, and must be linked to the corresponding dataset record (data card) in the data catalogue. Conditional approvals must enumerate specific technical controls, scope limitations, or source exclusions that constitute binding requirements for the use of the corpus — for example, restriction to non-commercial research, exclusion of specific article types, or mandatory redaction of personal health information. Conditional requirements must be treated as enforceable obligations and must be verified as part of any ongoing monitoring and periodic re-assessment process.

8.2.7 Escalation pathways and governance committee authority

Ingestion decisions involving high-risk AI systems, novel or legally uncertain bases, sensitive domains, or significant unresolved ambiguity must be escalated to the AI governance committee or designated authority for binding resolution, rather than decided at project level. Escalation pathways and the thresholds that trigger escalation must be defined in advance within the governance policy, ensuring that project teams have clear, unambiguous guidance on when independent authority is required. The existence of explicit escalation thresholds is itself a governance requirement under the NIST AI RMF Govern function, as it ensures that risk decisions are not systematically resolved at the level of least organisational accountability.

8.2.8 Ongoing monitoring and re-assessment triggers

Approved ingestion decisions are not static. They must be subject to periodic re-assessment and must be automatically triggered for review upon: changes to applicable law or regulatory guidance; publisher policy updates or newly issued machine-readable TDM opt-out notices; changes to the AI system's deployment scope or commercialisation status; and rights-holder objections or third-party complaints. Data-catalogue tooling must implement automated alerts linked to source-registry entries, prompting governance review when licensing terms are updated or when new opt-out notices are detected for already-approved corpora.

Ongoing monitoring must be assigned as a standing responsibility to named data owners and the data governance office, with escalation to legal and IP counsel where substantive changes in the regulatory or licensing environment are identified. The monitoring function must be adequately resourced and must operate independently of the incentive structures of the AI development team. This separation of responsibilities is essential to maintaining the integrity of the governance process over the full lifecycle of the AI system, including any post-deployment updates that expand the scope or nature of the corpus in use.

8.3 Audit trail and evidence management

Robust audit trail and evidence management constitutes a foundational pillar of AI governance in research contexts where scientific literature is used as training data. In the specific setting of

large language model (LLM) development incorporating medical and biomedical corpora, the governance imperative extends beyond internal quality assurance to encompass regulatory compliance under instruments such as the EU AI Act, the General Data Protection Regulation (GDPR), and the text and data mining (TDM) exception regime established under Directive (EU) 2019/790 on copyright in the Digital Single Market (CDSM Directive). This section sets out the structural requirements and operational mechanisms through which the project ensures comprehensive traceability of data assets, decision records, and model artefacts throughout the full lifecycle of AI system development.

8.3.1 Data Catalog Infrastructure

Data Catalog Architecture and Lineage Tracking

The governance architecture for literature-derived training corpora is anchored in a centralised, machine-readable data catalog that functions as the primary authoritative registry of all datasets ingested and processed within the project. Each corpus derived from scientific literature sources—whether obtained via publisher APIs, open-access repository dumps, OAI-PMH harvests, institutional subscription services, or web-based TDM crawls—is registered as a discrete, autonomous catalogued asset. The catalog entry for each corpus captures, at minimum, its provenance metadata (source type, access mechanism, ingestion timestamp), legal status (applicable legal basis, licensing terms, TDM opt-out status), data quality indicators, and linkage to any downstream model artefacts for which the corpus was used as training or evaluation input. By consolidating these dimensions within a single authoritative record, the catalog provides the operational backbone for compliance verification and inter-departmental governance review.

End-to-end data lineage is established and maintained through automated instrumentation of the data pipeline. Lineage records are required to trace the complete transformation chain from raw source ingestion through all intermediate processing stages—including deduplication, filtering, de-identification, format normalisation, and quality validation—to the final training and evaluation splits produced for each model version. Each transformation operation is captured as a discrete, timestamped lineage node, enabling post-hoc forensic reconstruction of which specific articles, document versions, or corpus subsets contributed to any given model iteration. This granularity is operationally essential for incident response scenarios—for instance, where a rights holder invokes legal remedies relating to unauthorised reproduction—as well as for regulatory inspection under Article 53 of the EU AI Act, which requires providers of general-purpose AI models to maintain technical documentation on training data.

Corpus and Metadata Versioning

Every training corpus is managed under a formal versioning scheme that associates each model training run with an immutable, uniquely identified snapshot of the underlying dataset and its associated metadata. Versioning serves two distinct governance objectives. First, it enables reproducibility of training results and supports the retrospective assessment of how incremental dataset changes—such as the addition of new source categories, the removal of retracted publications, or the application of updated de-identification procedures—affect model behaviour and evaluation performance. Second, and critically from a legal compliance standpoint, metadata versioning must extend explicitly to legal and licensing records, capturing the state of any applicable TDM opt-out declarations and access rights at the precise moment of corpus ingestion. This is particularly relevant under Article 4(3) of the CDSM Directive, where

rights holders may declare an opt-out against TDM use of their works, and where the temporal boundary of that declaration affects the legal basis available to the research organisation. Immutable versioning of the legal metadata associated with each corpus snapshot provides the evidentiary foundation necessary to demonstrate that opt-out declarations were respected in accordance with the timing and mechanism requirements of the applicable framework.

Centralised Inventory, Ownership, and Usage-Context Records

The project maintains a unified inventory that registers all literature-derived corpora as independently managed governance assets, independent of the particular AI system or research workstream that first produced them. The inventory enables governance bodies to assess cumulative legal exposure across concurrent or sequential projects, identify overlapping or duplicated corpora that may carry redundant licensing obligations, and apply consistent stewardship standards irrespective of the originating team. Each catalogued dataset record must declare a named data steward—an individual or organisational unit—with defined accountability for ongoing legal compliance monitoring, authorisation of reuse decisions, and escalation of any changes in source licensing status to the governance committee.

Dataset catalog records embed direct cross-references to the corresponding legal and intellectual property (IP) assessment documents, privacy review outcomes, and any conditional approvals issued by the governance committee or legal counsel. This linkage ensures that governance decisions remain traceable to specific dataset versions and cannot become decoupled from the technical records over time. Usage-context metadata further specifies the AI systems, use cases, and deployment environments for which a given corpus is authorised, thereby preventing inadvertent reuse of a research-exemption corpus in a commercial deployment context without the mandatory re-evaluation that such a change in purpose would necessitate under both copyright and data protection law.

8.3.2 Dataset and Model Documentation

Data Cards

Structured dataset documentation, operationalised in the form of data cards, provides standardised, human-readable summaries of corpus origins, collection methodology, intended purpose, ethical considerations, and known technical limitations. In the context of literature-based AI development, data cards function as governance instruments rather than merely informational artefacts: their completion is a mandatory condition for a corpus to be promoted from a staging environment to production use in model training. Data card templates are aligned with the technical documentation obligations under the EU AI Act—including, where applicable, the requirements introduced by the Act's provisions concerning general-purpose AI systems with respect to copyright compliance summaries—so that regulatory disclosure requirements are embedded in standard operational workflows rather than treated as post-hoc reporting tasks.

Each data card must enumerate all source categories contributing to the corpus, including peer-reviewed journal APIs, open-access repositories such as PubMed Central, preprint servers, institutional repositories, and licensed aggregators, together with the specific technical access mechanism used for each. This enumeration is a prerequisite for accurate legal-basis assignment, since the applicable copyright and TDM normative framework may differ across access mechanisms even when the underlying content is materially identical—for instance, a publication accessible via an open-access repository under a Creative Commons licence may

simultaneously be accessible via a subscription API under contractual terms that impose different TDM restrictions. The data card must also formalise the selection logic applied to the corpus, documenting inclusion parameters—such as publication year range, language, disciplinary scope, and document type—as well as exclusion rules covering retracted articles, publications indexed in predatory journal registries, and rights-reserved sources for which no applicable legal basis has been established.

Model Cards

Model documentation, maintained in the form of model cards, provides standardised records encompassing intended use cases, quantitative performance metrics, references to the specific dataset versions used in training and evaluation, known limitations, and ethical considerations relevant to the deployment context. Model card maintenance is continuous throughout the model lifecycle: each retraining, fine-tuning, or incremental update event triggers a mandatory card revision that links the resulting model version to the corpus snapshot—identified by its immutable version identifier—used in that training run. Completion of an updated model card constitutes a formal process control gating deployment, operationalising documentation as a compliance mechanism rather than a discretionary transparency measure.

Legal Bases, Opt-Out Compliance, and Sensitive Data Disclosures

Each data source within a corpus must be assigned a documented legal basis for its use in AI training. The applicable normative categories, which must be selected and recorded explicitly, include: use under an open licence (e.g., Creative Commons Attribution); use under an institutional subscription agreement that expressly permits TDM; use pursuant to the research TDM statutory exception under Article 3 of the CDSM Directive; or use under a separate contractual authorisation. Where rights holders have exercised a machine-readable TDM opt-out—as permitted under Article 4(3) of the CDSM Directive and increasingly implemented via standards such as the Robots Exclusion Protocol or publisher-specific API flags—the dataset record must flag the affected sources as excluded, with the opt-out detection date, detection mechanism, and the pipeline stage at which exclusion was applied, all recorded to demonstrate that the project exercised due diligence commensurate with the obligations under applicable law. Legal-basis records are subject to periodic review and must trigger re-assessment automatically when source licensing terms change or when the intended use of a corpus migrates from a research context to a commercial deployment environment.

Where a corpus contains personal or sensitive data—including patient narratives, author identifiers, genomic sequences, digital pathology images, or clinical case descriptions—the data card must disclose the relevant data categories, the applicable legal basis for processing under GDPR or equivalent regulation, and the safeguards applied. Applied de-identification, pseudonymisation, aggregation, or redaction operations are documented as discrete processing steps within the lineage record, specifying the method, tool, and validation approach employed. Where a Data Protection Impact Assessment (DPIA) was conducted in accordance with Article 35 GDPR, the assessment reference and outcome are linked to the dataset record and subject to review upon any material change to corpus scope or processing purpose.

Dataset documentation must further include a structured bias and coverage analysis that identifies systematic over- or under-representation within the corpus across dimensions including language, geographic region, research discipline, publisher ownership, and study design type. Coverage gaps and identified biases are propagated as flagged risk factors in the associated model card, with corresponding mitigating controls—such as targeted corpus

expansion or the construction of domain-adapted evaluation sets—specified and tracked within the governance register.

8.3.3 Logging, Audit Trails, and Access Control

Access and Processing Logs

Comprehensive access logging is required for all interactions with literature-derived corpora. Access logs must record, with sufficient granularity to support security monitoring, incident investigation, and regulatory audit, all retrieval and export operations—including bulk export events and queries directed at corpora containing sensitive topic areas such as clinical data or genomic information. Log retention periods are aligned with applicable legal requirements and with the operational lifecycle of the models trained on the relevant corpus, ensuring that audit trails remain available for the full duration of potential liability exposure under applicable copyright, data protection, and sector-specific regulatory instruments.

A separate processing log captures each pipeline operation applied to the corpus from ingestion through to final dataset assembly. This encompasses ingestion job parameters, filtering rules applied and their outcomes, de-identification and redaction steps, quality-validation check results, and any manual curation interventions including the identity of the approving personnel and the basis for the decision. Processing logs constitute the primary technical evidence base for demonstrating regulatory compliance in the event of inspection: they must be sufficiently detailed to allow an external auditor or regulatory authority to reconstruct the data preparation chain and verify that all required safeguards were applied at the appropriate pipeline stages.

8.3.4 Training and Evaluation Logs

Training logs establish a verifiable, machine-readable mapping between dataset version identifiers, model version identifiers, hyperparameter configurations, computational environment specifications, and evaluation results. This traceability chain enables the project to demonstrate, for any deployed model version, precisely which corpus content—including which specific document versions and which licensed or exempted sources—informed the model's parameters. Evaluation logs must record the specific literature corpora used in benchmark construction, enabling systematic detection of dataset contamination resulting from train-test overlap, a methodological risk of particular significance in biomedical domains where the pool of high-quality annotated corpora is constrained.

Access-Control Records

A formal access-control framework defines and enforces differentiated permissions for the three critical corpus operations: ingestion of new source material into the catalog, modification of existing corpus versions or their associated metadata, and export for model training or external sharing. Each permission tier is governed by a defined authorisation basis—role assignment, project-level approval, or governance committee decision—and access-control decisions are logged with reference to that basis, creating a full auditable record of which individuals or systems held which permissions at each point in the corpus lifecycle. For corpora containing sensitive content—human genomic data, digital pathology image series, or identified clinical case records—access is subject to additional restrictive controls and periodic entitlement review cycles consistent with the data stewardship norms applicable in high-data-intensity biomedical research environments, including those imposed by institutional review boards and data access committees.

8.3.5 Documentation Quality and Governance Metrics

Documentation Quality Audits

Periodic structured audits of data cards and model cards assess completeness, accuracy, and consistency against defined templates and applicable regulatory requirements. Documentation quality is treated as an auditable governance control in its own right: the audit process examines not only whether fields are populated, but whether the content is internally consistent, accurately references the corresponding legal assessments, and satisfies the specificity required for regulatory scrutiny. Audit findings are recorded in the governance register and escalated to the AI governance committee where identified gaps indicate material compliance risk. Remediation timelines are formally tracked, and unresolved gaps may result in the suspension of corpus promotion or model deployment pending resolution, ensuring that documentation obligations carry concrete operational consequences rather than functioning as a symbolic compliance exercise.

Governance Metrics, Readiness Scores, and Residual Risk Ratings

A defined set of quantifiable governance metrics is maintained for each literature-derived corpus, covering dimensions including legal-basis completeness, lineage coverage rate, documentation readiness, de-identification verification status, and opt-out compliance rate. These metrics are aggregated into a composite readiness score that provides governance bodies with a standardised signal for deployment decisions, enabling risk-proportionate oversight without requiring full committee review of every incremental corpus update. Readiness scores are versioned alongside the corpus and made available to downstream consumers of the dataset through the catalog record.

Following each governance review cycle, a residual risk rating is assigned to each corpus, reflecting outstanding legal, ethical, or technical risks that remain after the application of documented controls. The residual risk rating is propagated to the associated model card, ensuring that deployment teams and subsequent governance reviewers are informed of the risk profile of the underlying training data. Where the residual risk rating exceeds a defined threshold, escalation to senior governance oversight is mandatory prior to deployment authorisation. This mechanism operationalises the principle that AI system accountability must be traceable, through the full documentation chain, from deployed model behaviour to the specific data governance decisions that shaped the model's training.

9 Risk Assessment and mitigation measures

The use of medical scientific literature for LLM training gives rise to a combination of legal, regulatory, ethical, reputational, and operational risks that cannot be addressed solely through abstract legal interpretation. These risks emerge at different points of the data lifecycle and may affect not only the lawfulness of source use, but also the traceability, defensibility, and overall reliability of the consortium's data-processing practices. For this reason, the present section adopts a structured risk-oriented perspective. It identifies the principal categories of risk associated with the collection, processing, and reuse of medical literature in the project context, and links each category to corresponding mitigation measures that should be embedded into the consortium's compliance and governance workflow. The next tables provide a consolidated

overview of these risks and serve as a practical reference framework for the more detailed discussion developed in the following subsections.

Table 8. Legal, data protection, and governance risks related to the use of medical literature for LLM training

Risk category	Example risk	Potential impact	Main mitigation measure
Legal and IP	Reliance on a TDM exception where the legal conditions are not actually met	Copyright infringement claims, invalid source use, and possible removal of materials from the corpus	Conduct source-by-source eligibility assessment and document the legal basis before ingestion
Legal and IP	Failure to identify a valid Article 4 opt-out or applicable licence restriction	Unauthorised use of protected material and exposure to contractual or rights-based claims	Implement opt-out and licence screening as a mandatory pre-ingestion control
Legal and IP	Breach of subscription or database access conditions through automated extraction	Contractual breach, access loss, reputational damage, and interruption of processing activities	Review access terms in advance and restrict ingestion methods to authorised channels
Data protection and confidentiality	Inclusion of personal or health-related data without a valid legal basis or adequate safeguards	GDPR non-compliance, regulatory exposure, and heightened project risk	Assess whether personal data are present, identify the legal basis, and apply minimisation, safeguards, and documentation measures
Data protection and confidentiality	Model memorisation or extraction of personal information from training data	Privacy breach, regulatory scrutiny, and reputational harm	Limit high-risk inputs and assess technical controls against memorisation and inference risks
Regulatory and governance	Inadequate documentation or traceability of source-selection decisions	Inability to justify compliance during audit, review, or dispute	Maintain a complete audit trail covering provenance, legal assessment, decisions, and restrictions
Regulatory and governance	Misalignment between dataset practices and AI Act-related governance expectations	Weak defensibility of downstream AI development and increased compliance burden later in the project	Embed governance, documentation, and data-quality controls early in the data lifecycle

Table 9. Ethical, reputational, and operational risks related to the use of medical literature for LLM training

Risk category	Example risk	Potential impact	Main mitigation measure
Ethical and scientific quality	Non-representative or biased corpus composition affecting downstream model behaviour	Reduced reliability, unfair outputs, weak scientific validity, and possible regulatory concern	Apply source-diversity, representativeness, and review criteria during corpus design

Reputational	Use of controversial, disputed, or poorly documented sources	Loss of trust among partners, reviewers, or external stakeholders	Escalate ambiguous or high-impact cases and record the rationale for final decisions
Operational	Late identification of restricted sources after ingestion or preprocessing	Rework, delays, resource loss, and possible need to rebuild parts of the corpus	Front-load legal screening and introduce review checkpoints before each major lifecycle stage
Operational	Inconsistent application of acceptance and rejection criteria across partners	Uneven compliance standards and reduced defensibility of consortium decisions	Apply shared decision criteria, governance rules, and common documentation templates
Operational	Insufficient coordination between legal review and technical handling	Legally acceptable sources may be processed in technically non-compliant ways, or vice versa	Ensure coordinated review between legal/IP, technical, and WP2 governance functions
Reputational and governance	Escalation thresholds are unclear or applied too late	Sensitive issues may remain unresolved until they affect the wider project	Define escalation triggers in advance and route medium- and high-risk cases through the formal governance workflow

9.1 Legal and IP risks

This section identifies and analyses the principal legal and intellectual property (IP) risks arising from the use of medical scientific literature in the training of large language models (LLMs). The analysis addresses copyright and database-right infringement, licence-term compliance, contractual restrictions, regulatory obligations under the EU copyright and AI legislative framework, and data-protection requirements specific to biomedical corpora. Each risk dimension is examined with reference to the applicable legal instruments and their governance implications for the project.

9.1.1 Copyright Infringement

The unauthorised reproduction of protected scientific publications during the construction of an AI training corpus constitutes copyright infringement irrespective of whether the resulting model is deployed commercially or for research purposes. The critical legal act is the reproduction itself: the ingestion of copyrighted content into a training dataset triggers reproduction rights at the moment of copying, and the downstream transformation effected by the model's training process does not retroactively cure the initial unlawful reproduction. This distinction is legally significant because it displaces arguments premised solely on the transformative nature of machine learning.

The financial exposure associated with copyright infringement at scale is substantial. Under United States law, for example, statutory damages may reach up to USD 150,000 per work in cases of wilful infringement, a figure that compounds rapidly when applied to large corpora comprising thousands of scientific articles. Equivalent provisions exist under national implementations of EU Directive 2001/29/EC. Accordingly, the volume and provenance of training data are not merely technical parameters but determinants of material legal liability.

9.1.2 Database-Right Infringement

Independently of the copyright subsisting in individual scientific articles, scientific literature aggregators and publisher databases may hold sui generis database rights under EU Directive 96/9/EC on the legal protection of databases. These rights protect the substantial investment made in the obtaining, verification, or presentation of database contents, and prohibit the extraction or re-utilisation of substantial portions without authorisation. Crucially, database rights operate as a distinct legal layer: bulk harvesting of records from licensed platforms such as Web of Science or PubMed aggregators may infringe database rights even where the underlying article copyright is separately licensed or has expired.

The assessment of legality must therefore proceed on two parallel tracks, evaluating both the copyright status of individual publications and the database-right position of each aggregated source. A finding of lawful access at the article level does not, of itself, establish a lawful basis for systematic extraction from the hosting platform.

9.1.3 Use of Pirated or Unlawfully Obtained Content

Sourcing scientific articles from unauthorised repositories — commonly referred to as shadow libraries — constitutes copyright infringement at the point of download, regardless of any subsequent legitimate use applied to the material or any fair-use or text-and-data mining (TDM) argument advanced in relation to the training process itself. The infringing act is complete upon acquisition, and the lawfulness of subsequent processing steps cannot retroactively extinguish liability arising from unlawful access.

Of particular governance significance is the principle that remedial acquisition of lawful copies after an unlawful download does not cure the original infringement. This renders post-hoc curation strategies legally ineffective as a primary risk-mitigation mechanism. The appropriate control is preventive: data governance frameworks must require a documented, auditable lawful-access pathway for each source corpus as a mandatory prerequisite for ingestion approval, thereby eliminating exposure at the point of origin rather than attempting remediation after the fact.

9.1.4 Licence-Term Violations

A significant proportion of scientific literature is published under open-access or Creative Commons licences that, while permitting access and redistribution, impose conditions whose breach converts an otherwise lawful access scenario into an act of infringement. Attribution requirements, non-commercial use restrictions, and share-alike obligations are among the most common such conditions. Non-compliance with any of these terms may vitiate the licence entirely, exposing the processing entity to the full range of copyright remedies as though no licence had been granted.

Beyond open-access licences, institutional subscription agreements frequently contain explicit restrictions on automated bulk downloading, systematic text mining, and the use of retrieved content for AI training purposes. Non-compliance with these contractual provisions gives rise to a dual exposure: liability in contract vis-à-vis the publisher and, depending on the jurisdiction and the nature of the restriction, potential copyright infringement where the contractual limitation mirrors or reinforces a copyright-protected exclusive right. Systematic licence-condition tracking, applied at the level of individual articles or defined corpora, is therefore a governance imperative rather than an administrative convenience.

9.1.5 Breach of Publisher Contractual Restrictions

In addition to the licence-term risks addressed above, publisher API terms and platform terms of service frequently impose restrictions on AI training uses that operate independently of underlying copyright law. Content may be technically accessible via an API or a licensed platform, yet subject to contractual prohibitions on use for model training. These restrictions constitute a separate layer of contractual risk that cannot be displaced by copyright exceptions or TDM provisions, as they arise from freedom-of-contract principles rather than from the statutory framework.

Contract review must accordingly be embedded as a mandatory step in the data-intake workflow, with legal and IP counsel providing binding clearance before any new publisher source is onboarded into a training corpus. The practice of deferring contract review until after data collection has commenced is not operationally acceptable from a risk-management perspective, as it creates an interval of unquantified contractual exposure.

9.1.6 Violation of TDM Opt-Out and Reservation-of-Rights Mechanisms

Under Article 4 of EU Directive 2019/790 on copyright in the Digital Single Market (CDSM Directive), rights holders may reserve their rights against commercial text-and-data mining by means of machine-readable opt-out notices. The effect of a valid opt-out is to remove the content from the scope of the commercial TDM exception provided in Article 4, leaving no statutory basis for the processing. Proceeding with ingestion without checking and respecting such notices therefore constitutes regulatory non-compliance and, simultaneously, an act of copyright infringement for which no safe harbour is available.

Opt-out mechanisms are not static: publisher policies evolve, and rights reservations may be introduced, modified, or withdrawn over time. AI governance workflows must therefore incorporate not only an initial automated or systematic check for machine-readable opt-out flags prior to ingestion, but also a periodic re-assessment regime triggered whenever publisher policies change or when corpora are refreshed. The legal consequence of disregarding a valid opt-out extends beyond infringement liability: it vitiates reliance on the TDM exception as a legal basis for the entire processing activity, removing the primary safe harbour available under EU law for commercial AI training operations.

9.1.7 Regulatory Non-Compliance under the EU AI Act and CDSM Framework

The EU AI Act (Regulation (EU) 2024/1689) imposes affirmative obligations on providers of general-purpose AI (GPAI) models that extend beyond those arising under copyright law. Specifically, providers are required to publish a sufficiently detailed summary of the copyrighted content used in training and to demonstrate compliance with applicable TDM and copyright rules. IP governance thereby becomes a direct regulatory obligation rather than a purely contractual or tortious matter, and non-compliance may attract regulatory sanctions in addition to civil copyright liability.

The failure to maintain and disclose adequate training-data documentation constitutes a standalone violation of the AI Act's transparency requirements, separate from and cumulative with any underlying copyright infringement. This creates a layered compliance structure: lawful access and respect for opt-out mechanisms are prerequisites for invoking the TDM exception under the CDSM Directive, which in turn constitutes a prerequisite for satisfying GPAI regulatory compliance under the AI Act. A deficiency at any point in this chain propagates upward, rendering the entire compliance position untenable. The interaction between these two

instruments demands that IP governance, copyright clearance, and regulatory documentation procedures be treated as an integrated compliance system rather than discrete workstreams.

9.1.8 Data-Protection Obligations Arising from Personal Data in Biomedical Literature

Scientific publications in clinical, genomic, and biomedical domains frequently contain personal or sensitive personal data — including patient narratives, identifiable case reports, and genomic sequences — embedded within the text. The ingestion of such content into an AI training pipeline constitutes processing of personal data within the meaning of the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) and triggers the full range of data-protection obligations thereunder, irrespective of any copyright clearance that may have been obtained in relation to the publication.

Sensitive categories of data as defined in GDPR Article 9 — which encompass health data, genetic data, and other special categories — attract a heightened protection regime. Processing such data requires not only an identified lawful basis under Article 6 but also a derogation under Article 9(2) and, in most cases, additional technical and organisational safeguards beyond those applicable to ordinary personal data. Any processing of personal data present in scientific literature must therefore be grounded in a corpus-specific and domain-aware legal-basis analysis, accounting for sector-specific ethics frameworks such as biomedical research ethics and human-subjects regulation that may constrain permissible uses even where a general GDPR basis is otherwise available.

High-risk processing scenarios — including large-scale processing of health or genetic data for AI model training — typically trigger the obligation to conduct a Data Protection Impact Assessment (DPIA) pursuant to GDPR Article 35 prior to commencing processing. Omitting this assessment constitutes a procedural violation independent of any substantive data-protection harm and may attract supervisory authority enforcement action. Technical safeguards such as de-identification, pseudonymisation, aggregation, and processing within access-restricted environments must be implemented as governance controls, not as optional enhancements, wherever personal or sensitive data is present in the training corpus.

9.1.9 Insufficient Documentation and Traceability as a Systemic Governance Failure

The absence of data-lineage records, dataset documentation (data cards), and model cards constitutes a systemic governance deficit that undermines both internal accountability and the ability to demonstrate regulatory compliance to supervisory authorities or courts. The evidentiary function of documentation is heightened in the AI training context because the information required to reconstruct the provenance of a model — which corpora were used, in which versions, under which access pathways, and subject to which rights-clearance assessments — is not inherently preserved by the model itself and must be captured through deliberate governance controls applied during corpus construction and model development.

The EU AI Act's technical-documentation and transparency requirements effectively mandate traceability of training data provenance for GPAI models. An organisation that cannot reconstruct the corpora, versions, and access pathways used in a given model version cannot satisfy these obligations, regardless of whether the underlying IP and data-protection controls were in fact compliant at the time of ingestion. Documentation must therefore be operationalised as a governance gate — integrated into ingestion pipelines and deployment workflows — rather than treated as a post-hoc reporting artefact. Only when documentation functions as a prospective control, embedded in operational processes from the outset, can it serve its dual purpose as

both a genuine compliance instrument and an accountability mechanism capable of withstanding regulatory scrutiny.

9.2 Ethical, regulatory, and reputational risks

The integration of scientific literature into large language model (LLM) training pipelines gives rise to a multidimensional risk landscape that spans ethical, scientific integrity, governance, regulatory, and reputational dimensions. Each category of risk is structurally distinct, yet the categories are mutually reinforcing: deficiencies in governance operationalization, for instance, compound both regulatory exposure and reputational vulnerability. This section systematically characterises each risk category in the context of AI systems trained on medical and scientific corpora, identifying the specific mechanisms of harm and the governance controls required to mitigate them.

9.2.1 Ethical Risks

Bias and Representation Risk in Corpus Composition

Systematic imbalance in the composition of training corpora constitutes a structural data-quality defect with direct ethical implications. When scientific literature datasets overrepresent particular geographic regions, publication languages, or research disciplines, the resulting models reflect those asymmetries in their outputs, regardless of the technical sophistication applied during training. Such imbalance is not a stochastic artefact but a foreseeable and documentable phenomenon, arising from the uneven global distribution of indexed scientific production and from the selective access policies of major academic publishers. Corpus composition bias must therefore be identified, assessed, and mitigated as part of the data-intake and curation stage, and not treated as a post-deployment correction problem. Appropriate governance controls at the ingestion phase include diversity audits of candidate datasets, stratified sampling procedures, and documented justifications for any systematic exclusion of source categories.

Amplification of Inequities in Research Attention and Clinical Evidence

Beyond the structural composition of training corpora, there is a distinct and more consequential risk that AI systems encode and propagate existing disparities in research visibility and clinical evidence coverage. The global distribution of published biomedical research is markedly skewed toward high-income regions and majority populations; diseases disproportionately affecting low-income countries or minority clinical groups remain systematically underrepresented in the indexed literature. When AI models are trained on such corpora, they are liable to generate outputs that reflect and reinforce these disparities, producing recommendations, summaries, or decision-support outputs of systematically lower reliability for underrepresented populations. In clinical or biomedical deployment contexts, this amplification effect carries direct ethical consequences and may constitute a violation of applicable non-discrimination and equal-treatment obligations under AI governance frameworks, including the risk management obligations of the EU AI Act for high-risk systems.

Misuse and Dual-Use Risk from Sensitive Methodological Content

Scientific literature in domains such as synthetic biology, virology, or pharmacology may contain detailed technical methodologies that, when recombined, synthesised, or surfaced at scale by AI systems, could enable applications with significant potential for harm. This dual-use risk is not

hypothetical: the convergence of accessible AI capabilities with granular methodological information from peer-reviewed sources represents a materially different risk profile compared to a researcher consulting individual papers in isolation. Effective governance requires domain-specific pre-ingestion assessment to identify sensitive content categories, followed by the application of purpose-built controls including exclusion filters, access restriction mechanisms, and documented misuse impact assessments. The adequacy of these controls must be reviewed periodically and updated in response to the evolving capabilities of deployed model versions.

9.2.2 Scientific Integrity Risks

Generation of Scientifically Inaccurate Outputs

LLMs trained on scientific corpora are capable of generating outputs that carry the surface characteristics of authoritative scientific discourse while containing material factual inaccuracies. This failure mode arises from the statistical nature of language model inference, which optimises for linguistic plausibility rather than epistemic accuracy. In applied contexts — particularly where outputs are consumed by non-specialist users or integrated into decision-support workflows — the apparent authority of AI-generated scientific content may suppress appropriate critical scrutiny. Scientific integrity risk is accordingly a governance concern requiring dedicated control mechanisms: structured validation workflows, systematic output monitoring, and explicit user disclosures regarding the nature and limitations of model-generated content. The absence of such controls constitutes a foreseeable governance gap with direct consequences for the reliability of AI-assisted knowledge systems.

Hallucinated and Fabricated Bibliographic References

A specific and particularly damaging instance of the broader scientific integrity risk is the tendency of language models to produce plausible but non-existent bibliographic references. Citation hallucination — the generation of syntactically and stylistically credible citations that do not correspond to any real publication — represents a direct threat to research integrity and to the credibility of any system deployed in academic or clinical information contexts. The consequences range from the propagation of unfounded claims in the scientific record to institutional reputational damage for organisations relying on AI-assisted literature review. Mitigation requires a combination of technical controls, such as retrieval-augmented generation architectures and automated citation verification pipelines, and governance measures including mandatory human-in-the-loop review for high-stakes outputs and disclosure obligations to end-users regarding the verification status of AI-generated references.

Misrepresentation of Evidential Weight and Scientific Consensus

AI systems trained on scientific literature are at risk of distorting the evidential weight, consensus status, or generalisability of research findings. This may manifest as the unwarranted elevation of preliminary or contested findings to the status of established fact, the conflation of correlation with causation, or the application of population-level evidence to clinical contexts for which it is insufficiently validated. Such misrepresentation is particularly hazardous in medical AI applications, where the framing of evidence directly influences clinical decision-making. Governance controls must include mechanisms to prevent or flag unwarranted generalisation, and model documentation — including model cards and system disclosures — must explicitly characterise known limitations regarding evidential accuracy and scope of application.

9.2.3 Governance Operationalization Risks

Lack of Transparency in Training Data Documentation

Incomplete or absent documentation of training corpora, data provenance, applicable legal bases, and preprocessing decisions constitutes a transparency failure with regulatory and governance consequences. Under the EU AI Act, providers of AI systems — particularly those classified as high-risk — are subject to explicit technical documentation and transparency obligations that extend to the characteristics and curation methodology of training datasets. Similarly, ISO/IEC 42001 establishes data governance requirements that presuppose substantive documentation practices. Compliance with these frameworks cannot be achieved through symbolic documentation artefacts: data cards, model cards, and dataset statements must contain the operational detail necessary to permit meaningful audit, verification, and stakeholder scrutiny. Organisations that treat transparency obligations as a post-hoc formality, rather than as a design requirement embedded in the data pipeline, face exposure to regulatory enforcement as well as a loss of institutional credibility.

Insufficiency of Human Oversight Mechanisms

The absence of meaningful human-in-the-loop validation at critical stages of the AI development lifecycle — including dataset curation, output review, and deployment decision gates — represents a systemic governance gap. This gap is particularly significant in high-stakes application domains such as clinical decision support, where the consequences of inadequate oversight extend to patient safety. Human oversight must be formalised as a procedural requirement with defined review checkpoints, clear role assignments, and documented escalation criteria, rather than treated as a residual safeguard applied at the discretion of individual practitioners. The governance framework must specify, at minimum, the conditions under which human review is mandatory, the competencies required of reviewing personnel, and the remediation procedures applicable where review identifies outputs of insufficient quality or reliability.

Gap Between Governance Principles and Operational Implementation

Academic and regulatory reviews of AI governance practice consistently identify a structural gap between the articulation of governance principles and their concrete implementation at the organisational level. Governance frameworks that remain at the level of policy statements — without translation into enforceable procedures, named accountability assignments, and measurable compliance criteria — provide neither operational guidance nor meaningful risk control. Effective operationalization requires the deployment of structural accountability mechanisms: RACI matrices assigning ownership of governance functions, process gates conditioning dataset ingestion and model deployment approvals on documented compliance checks, and governance metrics enabling ongoing monitoring and audit. Without such mechanisms, governance frameworks function as reputational instruments rather than risk management tools.

9.2.4 Regulatory Risks

High-Risk Classification under the EU AI Act and Associated Obligations

Under Regulation (EU) 2024/1689 (the EU AI Act), AI systems intended for use in clinical decision-making or patient interaction are subject to high-risk classification, triggering a comprehensive set of pre-market and ongoing obligations. These include conformity assessment against harmonised standards, the preparation and maintenance of technical documentation, the implementation of quality management systems, and the establishment of post-market

monitoring arrangements. Transparency obligations require that end-users be provided with adequate information regarding system capabilities and limitations, and that human oversight mechanisms be embedded in the deployment context. Critically, the applicable risk classification is determined by the intended purpose of the system, and a material change in use case — such as a transition from research-only application to clinical deployment — constitutes a triggering event requiring re-assessment of all applicable regulatory obligations. Organisations that fail to conduct use-case classification prior to corpus ingestion and model deployment risk operating high-risk AI systems in a state of non-compliance, with attendant exposure to supervisory enforcement and civil liability.

9.2.5 Reputational Risks

Community Backlash over Perceived Appropriation of Scholarly Outputs

Even in cases where the training use of scientific literature is legally compliant — whether under applicable text and data mining exemptions or through the conclusion of licensing arrangements — the exploitation of scholarly outputs without meaningful author engagement, attribution mechanisms, or benefit-sharing arrangements may generate significant adverse reactions from scientific communities, rights-holder coalitions, and disciplinary associations. Community perception of appropriation is not bounded by legal compliance: stakeholders apply norms of scholarly credit, reciprocity, and consent that are distinct from, and often more demanding than, the minimum requirements of intellectual property law. Governance frameworks must therefore address community expectations as an independent risk variable, incorporating collaborative engagement processes, transparency obligations toward contributing authors and institutions, and clear communication of the scope and purpose of training data use. Failure to do so risks triggering organised opposition that may materially constrain future data access and collaborative partnerships.

Loss of Legitimacy with Scientific and Publishing Communities

Failure to align AI training practices with established disciplinary norms and community expectations carries the longer-term risk of institutional delegitimisation. Where researchers, publishers, and professional societies conclude that an organisation's practices are inconsistent with prevailing standards on attribution, scientific credit, and data stewardship, the consequences may include withdrawal of data-sharing arrangements, termination of publishing partnerships, and reputational damage that is difficult to remediate after the fact. Legitimacy in this context is an ongoing relational asset, not a one-time certification: it requires demonstrated adherence to community norms through transparent governance practices, proactive communication of applicable safeguards, and a documented process for responding to objections raised by researchers or rights-holders. Organisations that treat scientific community relations as peripheral to their AI governance strategy underestimate the operational dependencies these relationships represent.

Regulatory and Reputational Exposure from Unsanctioned Clinical Use of Outputs

The deployment of model outputs in clinical workflows without adequate validation, labelling, or oversight controls creates a dual exposure profile: regulatory liability under applicable health AI frameworks and reputational harm in the event of patient safety incidents. This risk category is particularly acute because clinical misuse may occur outside the direct control of the developing organisation, through the downstream adoption of outputs by clinical practitioners or healthcare institutions that have not been specifically authorised to use the system in that context. Risk controls must accordingly extend beyond technical output verification to include

governance-level restrictions on authorised use cases, mandatory disclosure mechanisms informing end-users of system limitations, and contractual or policy instruments capable of constraining deployment to contexts covered by the approved conformity assessment scope.

9.3 Mitigation measures

The integration of scientific literature into large language model (LLM) training pipelines gives rise to a multidimensional risk landscape that spans ethical, scientific integrity, governance, regulatory, and reputational dimensions. Each category of risk is structurally distinct, yet the categories are mutually reinforcing: deficiencies in governance operationalization, for instance, compound both regulatory exposure and reputational vulnerability. This section systematically characterises each risk category in the context of AI systems trained on medical and scientific corpora, identifying the specific mechanisms of harm and the governance controls required to mitigate them.

9.3.1 Ethical Risks

Bias and Representation Risk in Corpus Composition

Systematic imbalance in the composition of training corpora constitutes a structural data-quality defect with direct ethical implications. When scientific literature datasets overrepresent particular geographic regions, publication languages, or research disciplines, the resulting models reflect those asymmetries in their outputs, regardless of the technical sophistication applied during training. Such imbalance is not a stochastic artefact but a foreseeable and documentable phenomenon, arising from the uneven global distribution of indexed scientific production and from the selective access policies of major academic publishers. Corpus composition bias must therefore be identified, assessed, and mitigated as part of the data-intake and curation stage, and not treated as a post-deployment correction problem. Appropriate governance controls at the ingestion phase include diversity audits of candidate datasets, stratified sampling procedures, and documented justifications for any systematic exclusion of source categories.

Amplification of Inequities in Research Attention and Clinical Evidence

Beyond the structural composition of training corpora, there is a distinct and more consequential risk that AI systems encode and propagate existing disparities in research visibility and clinical evidence coverage. The global distribution of published biomedical research is markedly skewed toward high-income regions and majority populations; diseases disproportionately affecting low-income countries or minority clinical groups remain systematically underrepresented in the indexed literature. When AI models are trained on such corpora, they are liable to generate outputs that reflect and reinforce these disparities, producing recommendations, summaries, or decision-support outputs of systematically lower reliability for underrepresented populations. In clinical or biomedical deployment contexts, this amplification effect carries direct ethical consequences and may constitute a violation of applicable non-discrimination and equal-treatment obligations under AI governance frameworks, including the risk management obligations of the EU AI Act for high-risk systems.

Misuse and Dual-Use Risk from Sensitive Methodological Content

Scientific literature in domains such as synthetic biology, virology, or pharmacology may contain detailed technical methodologies that, when recombined, synthesised, or surfaced at scale by AI

systems, could enable applications with significant potential for harm. This dual-use risk is not hypothetical: the convergence of accessible AI capabilities with granular methodological information from peer-reviewed sources represents a materially different risk profile compared to a researcher consulting individual papers in isolation. Effective governance requires domain-specific pre-ingestion assessment to identify sensitive content categories, followed by the application of purpose-built controls including exclusion filters, access restriction mechanisms, and documented misuse impact assessments. The adequacy of these controls must be reviewed periodically and updated in response to the evolving capabilities of deployed model versions.

9.3.2 Scientific Integrity Risks

Generation of Scientifically Inaccurate Outputs

LLMs trained on scientific corpora are capable of generating outputs that carry the surface characteristics of authoritative scientific discourse while containing material factual inaccuracies. This failure mode arises from the statistical nature of language model inference, which optimises for linguistic plausibility rather than epistemic accuracy. In applied contexts — particularly where outputs are consumed by non-specialist users or integrated into decision-support workflows — the apparent authority of AI-generated scientific content may suppress appropriate critical scrutiny. Scientific integrity risk is accordingly a governance concern requiring dedicated control mechanisms: structured validation workflows, systematic output monitoring, and explicit user disclosures regarding the nature and limitations of model-generated content. The absence of such controls constitutes a foreseeable governance gap with direct consequences for the reliability of AI-assisted knowledge systems.

Hallucinated and Fabricated Bibliographic References

A specific and particularly damaging instance of the broader scientific integrity risk is the tendency of language models to produce plausible but non-existent bibliographic references. Citation hallucination — the generation of syntactically and stylistically credible citations that do not correspond to any real publication — represents a direct threat to research integrity and to the credibility of any system deployed in academic or clinical information contexts. The consequences range from the propagation of unfounded claims in the scientific record to institutional reputational damage for organisations relying on AI-assisted literature review. Mitigation requires a combination of technical controls, such as retrieval-augmented generation architectures and automated citation verification pipelines, and governance measures including mandatory human-in-the-loop review for high-stakes outputs and disclosure obligations to end-users regarding the verification status of AI-generated references.

Misrepresentation of Evidential Weight and Scientific Consensus

AI systems trained on scientific literature are at risk of distorting the evidential weight, consensus status, or generalisability of research findings. This may manifest as the unwarranted elevation of preliminary or contested findings to the status of established fact, the conflation of correlation with causation, or the application of population-level evidence to clinical contexts for which it is insufficiently validated. Such misrepresentation is particularly hazardous in medical AI applications, where the framing of evidence directly influences clinical decision-making. Governance controls must include mechanisms to prevent or flag unwarranted generalisation, and model documentation — including model cards and system disclosures — must explicitly characterise known limitations regarding evidential accuracy and scope of application.

9.3.3 Governance Operationalization Risks

Lack of Transparency in Training Data Documentation

Incomplete or absent documentation of training corpora, data provenance, applicable legal bases, and preprocessing decisions constitutes a transparency failure with regulatory and governance consequences. Under the EU AI Act, providers of AI systems — particularly those classified as high-risk — are subject to explicit technical documentation and transparency obligations that extend to the characteristics and curation methodology of training datasets. Similarly, ISO/IEC 42001 establishes data governance requirements that presuppose substantive documentation practices. Compliance with these frameworks cannot be achieved through symbolic documentation artefacts: data cards, model cards, and dataset statements must contain the operational detail necessary to permit meaningful audit, verification, and stakeholder scrutiny. Organisations that treat transparency obligations as a post-hoc formality, rather than as a design requirement embedded in the data pipeline, face exposure to regulatory enforcement as well as a loss of institutional credibility.

Insufficiency of Human Oversight Mechanisms

The absence of meaningful human-in-the-loop validation at critical stages of the AI development lifecycle — including dataset curation, output review, and deployment decision gates — represents a systemic governance gap. This gap is particularly significant in high-stakes application domains such as clinical decision support, where the consequences of inadequate oversight extend to patient safety. Human oversight must be formalised as a procedural requirement with defined review checkpoints, clear role assignments, and documented escalation criteria, rather than treated as a residual safeguard applied at the discretion of individual practitioners. The governance framework must specify, at minimum, the conditions under which human review is mandatory, the competencies required of reviewing personnel, and the remediation procedures applicable where review identifies outputs of insufficient quality or reliability.

Gap Between Governance Principles and Operational Implementation

Academic and regulatory reviews of AI governance practice consistently identify a structural gap between the articulation of governance principles and their concrete implementation at the organisational level. Governance frameworks that remain at the level of policy statements — without translation into enforceable procedures, named accountability assignments, and measurable compliance criteria — provide neither operational guidance nor meaningful risk control. Effective operationalization requires the deployment of structural accountability mechanisms: RACI matrices assigning ownership of governance functions, process gates conditioning dataset ingestion and model deployment approvals on documented compliance checks, and governance metrics enabling ongoing monitoring and audit. Without such mechanisms, governance frameworks function as reputational instruments rather than risk management tools.

9.3.4 Regulatory Risks

High-Risk Classification under the EU AI Act and Associated Obligations

Under Regulation (EU) 2024/1689 (the EU AI Act), AI systems intended for use in clinical decision-making or patient interaction are subject to high-risk classification, triggering a comprehensive set of pre-market and ongoing obligations. These include conformity assessment against harmonised standards, the preparation and maintenance of technical documentation, the implementation of quality management systems, and the establishment of post-market

monitoring arrangements. Transparency obligations require that end-users be provided with adequate information regarding system capabilities and limitations, and that human oversight mechanisms be embedded in the deployment context. Critically, the applicable risk classification is determined by the intended purpose of the system, and a material change in use case — such as a transition from research-only application to clinical deployment — constitutes a triggering event requiring re-assessment of all applicable regulatory obligations. Organisations that fail to conduct use-case classification prior to corpus ingestion and model deployment risk operating high-risk AI systems in a state of non-compliance, with attendant exposure to supervisory enforcement and civil liability.

9.3.5 Reputational Risks

Community Backlash over Perceived Appropriation of Scholarly Outputs

Even in cases where the training use of scientific literature is legally compliant — whether under applicable text and data mining exemptions or through the conclusion of licensing arrangements — the exploitation of scholarly outputs without meaningful author engagement, attribution mechanisms, or benefit-sharing arrangements may generate significant adverse reactions from scientific communities, rights-holder coalitions, and disciplinary associations. Community perception of appropriation is not bounded by legal compliance: stakeholders apply norms of scholarly credit, reciprocity, and consent that are distinct from, and often more demanding than, the minimum requirements of intellectual property law. Governance frameworks must therefore address community expectations as an independent risk variable, incorporating collaborative engagement processes, transparency obligations toward contributing authors and institutions, and clear communication of the scope and purpose of training data use. Failure to do so risks triggering organised opposition that may materially constrain future data access and collaborative partnerships.

Loss of Legitimacy with Scientific and Publishing Communities

Failure to align AI training practices with established disciplinary norms and community expectations carries the longer-term risk of institutional delegitimisation. Where researchers, publishers, and professional societies conclude that an organisation's practices are inconsistent with prevailing standards on attribution, scientific credit, and data stewardship, the consequences may include withdrawal of data-sharing arrangements, termination of publishing partnerships, and reputational damage that is difficult to remediate after the fact. Legitimacy in this context is an ongoing relational asset, not a one-time certification: it requires demonstrated adherence to community norms through transparent governance practices, proactive communication of applicable safeguards, and a documented process for responding to objections raised by researchers or rights-holders. Organisations that treat scientific community relations as peripheral to their AI governance strategy underestimate the operational dependencies these relationships represent.

Regulatory and Reputational Exposure from Unsanctioned Clinical Use of Outputs

The deployment of model outputs in clinical workflows without adequate validation, labelling, or oversight controls creates a dual exposure profile: regulatory liability under applicable health AI frameworks and reputational harm in the event of patient safety incidents. This risk category is particularly acute because clinical misuse may occur outside the direct control of the developing organisation, through the downstream adoption of outputs by clinical practitioners or healthcare institutions that have not been specifically authorised to use the system in that context. Risk controls must accordingly extend beyond technical output verification to include

governance-level restrictions on authorised use cases, mandatory disclosure mechanisms informing end-users of system limitations, and contractual or policy instruments capable of constraining deployment to contexts covered by the approved conformity assessment scope.

10 Best practices and compliance checklists

The legal and governance principles developed in the preceding sections must ultimately be translated into operational checks that can be applied consistently by the consortium during source selection, data handling, and preparation for AI use. The purpose of the present section is therefore not merely to restate the applicable obligations, but to convert them into practical control points that support repeatable and auditable decision-making. The following checklists are designed to function as working tools for the preliminary review of candidate sources, the verification of rights and restrictions, the assessment of privacy and governance safeguards, and the final validation of datasets before training or sharing activities take place. The next tables summarise these control points in a form intended to facilitate internal review, documentation, and compliance traceability across the project lifecycle.

Table 10. Source-selection and rights-clearance checklist prior to ingestion

Check item	Yes / No / N.A.	Evidence or note
Has the source been identified through a lawful and documented access route?		
Has the provenance of the source been recorded?		
Has the category of source been identified (e.g. article, abstract, figure, metadata, database record, supplementary material)?		
Has the rights status of the source been assessed?		
Has it been determined whether Article 3 or Article 4 CDSM may apply?		
If Article 4 is relevant, has a valid opt-out been checked and ruled out or documented?		
Have licence terms, subscription conditions, and platform restrictions been reviewed?		
Is the intended use compatible with the applicable licence or access terms?		
Has any database-right or platform-level extraction issue been considered?		
Has the preliminary outcome been recorded as accepted, rejected, or conditional / escalated?		

Table 11. Privacy, governance, and pre-training validation checklist

Check item	Yes / No / N.A.	Evidence or note
Has the presence or possible presence of personal data been assessed?		
If health-related or other special-category data may be present, has the applicable legal basis been identified?		
Have minimisation and safeguard requirements been considered and documented?		
Have any technical handling constraints been identified (e.g. segregation, restricted processing, controlled access)?		
Has the source or dataset been logged in the relevant compliance or traceability record?		
Has responsibility for approval, escalation, or monitoring been assigned?		
Has the source or dataset been reviewed for risks linked to representativeness, bias, or scientific suitability?		
Before training, has the dataset been reviewed to confirm that only approved and documented materials are included?		
Before sharing, have partner permissions, use restrictions, and onward-use constraints been checked?		
Has the final validation decision been recorded together with the supporting rationale?		

10.1 Best practices for source selection and lawful use

The construction of training corpora for large language models operating in the medical domain is subject to a complex and multi-layered legal framework that conditions every stage of the data acquisition and processing lifecycle. Compliance with this framework is not reducible to a single clearance decision at the point of publication; rather, it requires the systematic application of source-selection criteria, access verification procedures, licence analysis, and ongoing governance controls. This section sets out the principal best practices that must govern source selection and lawful use within any project seeking to train or fine-tune language models on scientific literature, with particular reference to the requirements of the EU Directive on Copyright in the Digital Single Market (CDSM Directive, 2019/790) and the EU AI Act.

10.1.1 Lawful Access as a Mandatory Prerequisite

The invocation of text and data mining (TDM) exceptions under Articles 3 and 4 of the CDSM Directive is predicated on the existence of lawful access to the works in question. Lawful access is not a procedural formality but a substantive threshold condition: content that is paywalled,

access-restricted, or otherwise unavailable without a valid subscription or equivalent access right cannot be processed under TDM exceptions in the absence of such entitlement. Accordingly, content eligibility for TDM processing must be established at the point of acquisition, with clear differentiation between open-access materials, subscription-based publications, and publicly available content. Any acquisition method that circumvents access controls — whether technical or contractual — forfeits the legal basis for downstream TDM processing and exposes the data controller to infringement liability.

Scientific corpora must therefore be acquired exclusively through publisher-sanctioned channels. These include institutional bulk-download facilities, dedicated publisher APIs, and authorised repository dumps. The access method is not merely a technical parameter: under EU law and most publisher licensing frameworks, the legal conditions applicable to TDM processing are tied to the manner in which the content was originally accessed. Rate limits, usage quotas, and platform-specific access instructions imposed by publishers and repository operators must be treated as binding constraints, both contractually and technically, rather than as operational preferences. Systematic non-compliance with these constraints constitutes a breach of the conditions under which lawful access was granted and may invalidate the TDM basis for any corpus built on such access.

10.1.2 Licence Analysis and Open-Access Compatibility

Where content is published under Creative Commons or equivalent open licensing regimes, a precise analysis of the applicable licence variant is required prior to corpus inclusion. Licences in the CC BY and CC BY-SA families are operationally compatible with AI training pipelines, provided that attribution obligations — which are mandatory under both variants — are addressed at the level of any publicly disseminated derived datasets or model documentation. The share-alike (SA) condition additionally requires that any publicly distributed adaptation or derived corpus be released under terms no more restrictive than the original licence, a requirement that has direct implications for dataset release strategies.

Content distributed under CC BY-NC or CC BY-NC-SA licences imposes a non-commercial use restriction that categorically precludes inclusion in commercial model training corpora unless an independent statutory authorisation — such as a research-oriented TDM exception or, in jurisdictions that recognise it, a fair use defence — independently establishes a sufficient legal basis. In EU member states implementing Article 3 of the CDSM Directive, the mandatory research TDM exception may provide such a basis, but only where the research purpose is genuine, non-commercial, and conducted by or in partnership with an eligible research organisation. Commercial deployment or redistribution of models trained under this exception is not authorised.

Works distributed under CC BY-ND or CC BY-NC-ND licences present a distinct category of risk arising from the NoDerivatives condition. Where the training process or the outputs of a trained model may be characterised as an adaptation or derivative work of the source material under applicable copyright law, inclusion of ND-licensed content in a training corpus may constitute licence infringement. Given the current absence of settled jurisprudence on whether and under what conditions LLM training constitutes the creation of a derivative work, the precautionary approach requires the systematic exclusion of ND-licensed works from training datasets intended for any form of commercial exploitation.

10.1.3 Verification of Publisher TDM Policies and Contractual Conditions

Publisher-specific TDM policies must be consulted and documented prior to corpus ingestion. Major scientific publishers have published formal TDM policies that specify whether, and under what conditions, authorised users — including institutional subscribers — may conduct TDM under the CDSM Directive framework. Some publishers additionally require the conclusion of specific TDM or AI-training licence agreements as a condition of lawful processing, even where statutory TDM exceptions would otherwise apply. The existence of such requirements must be verified on a publisher-by-publisher basis prior to the commencement of any corpus-building activity.

For commercial model development, reliance on research-oriented statutory exceptions is generally insufficient, as these exceptions are subject to opt-out mechanisms under Article 4 of the CDSM Directive and typically do not extend to the commercial deployment or redistribution of trained models. In such contexts, the negotiation of explicit bespoke TDM and AI-training licences with publishers is required. Publisher-imposed constraints — including prohibitions on the redistribution of full-text corpora, restrictions on model sharing, and limitations on the commercial use of TDM outputs — must be recorded in the project's dataset register and operationalised as enforceable controls within downstream product governance and data sharing arrangements.

10.1.4 Avoidance of Ad-Hoc Scraping and Ambiguous Provenance

Each source included in a training corpus must be assigned a documented and verifiable acquisition pathway — publisher API, authorised repository download, or institutional subscription access — since the applicable legal conditions for TDM vary by access mechanism and may differ materially from those governing human reading of the same content. Ad-hoc web scraping constitutes a high-risk acquisition method associated with uncertain provenance, unpredictable content quality, and a substantial likelihood of breaching platform terms of service, which may independently negate the TDM exception regardless of whether the content is otherwise freely accessible online. The practice of scraping should be avoided in favour of structured, sanctioned ingestion pipelines that preserve full provenance traceability from source to corpus entry.

10.1.5 Detection and Enforcement of Rightsholder Opt-Out

Under Article 4 of the CDSM Directive, rightsholders may reserve their rights against TDM for commercial purposes by means of machine-readable opt-out signals. The obligation to detect and respect such reservations is a compliance requirement that applies at the point of corpus construction. Automated tooling must therefore be implemented to identify rights-reservation indicators across sources, including machine-readable metadata embedded in publication records, publisher-level TDM policy declarations, and standard web crawling indicators such as robots.txt directives. This detection capability must be integrated into the corpus ingestion pipeline as a mandatory pre-processing gate, such that opted-out content is systematically excluded prior to any TDM processing.

Beyond opt-out detection, the corpus exclusion regime must be designed to address overlapping categories of ineligible content: works subject to a valid Article 4 opt-out, content distributed under NoDerivatives licence variants, and materials acquired in breach of platform terms of service. Exclusion rules across all three categories must be enforced through automated filtering mechanisms rather than reliance on manual review, both to achieve scalable compliance across

large corpora and to provide an auditable record demonstrating that exclusion obligations were systematically met.

10.1.6 Corpus Segmentation by Permitted-Use Scope

Content permissible solely under research-oriented TDM exceptions — including NC-licensed materials and works accessible only under Article 3 of the CDSM Directive — must be maintained in technically and logically distinct corpus partitions, segregated from content authorised for commercial model training. This segmentation is not merely an administrative preference; it constitutes a foundational data engineering control required to prevent the inadvertent inclusion of use-restricted content in commercial training runs, which would extinguish the statutory basis for its original inclusion. For organisations developing both research and commercial AI systems within shared infrastructure, corpus partitioning by permitted-use scope must be enforced at the storage, access control, and pipeline configuration layers.

10.1.7 Machine-Readable Provenance Registration and Dataset Documentation

A structured dataset register in machine-readable format must be established and maintained to capture provenance metadata for each element of the training corpus. The minimum required fields for each corpus entry include: the source URL or API endpoint from which the content was acquired; the acquisition date and method; the applicable legal basis for TDM processing, specifying whether this is a statutory TDM exception, a Creative Commons licence variant, a bespoke licence agreement, or public-domain status; and any opt-out or rights-reservation signals detected at the time of acquisition. Documentation must be maintained at corpus-level granularity to support dataset-specific legal review, enable targeted audit trails, and satisfy the public summary obligations under Article 53(1)(d) of the EU AI Act with respect to training data sources.

Machine-readable provenance records additionally serve as the technical substrate for integration with copyright auditing frameworks, including content fingerprinting and sampling techniques used to verify rightsholder due diligence. All filtering, transformation, and exclusion operations applied to the corpus during pre-processing must be logged as part of the provenance record, ensuring end-to-end traceability from raw source material to final training artefact. This traceability obligation must extend to derived artefacts — including intermediate caches, text embeddings, and fine-tuned model checkpoints — to support comprehensive copyright and privacy audits across the full AI development lifecycle. Provenance infrastructure must furthermore be designed to accommodate rightsholder takedown and opt-out requests, enabling the targeted removal of contested content from current and future training cycles in a verifiable and documented manner.

10.1.8 Sensitivity Tagging and Enhanced Review for High-Risk Sub-Corpora

Corpus segments with elevated sensitivity require identification and flagging prior to training in order to enable targeted application of enhanced data-protection and research ethics controls. In the context of medical scientific literature, relevant high-sensitivity categories include clinical journals containing detailed case reports, qualitative social science studies incorporating individual-level narratives, and any dataset elements that may contain or derive from personal health data as defined under Article 4(15) of the GDPR. Sensitivity tagging enables the targeted application of GDPR Article 9 special-category data safeguards, the initiation of Data Protection Impact Assessment (DPIA) procedures under Article 35, and compliance with sectoral health-data governance requirements, without imposing uniform overhead across the entirety of the corpus.

10.1.9 Ongoing Monitoring of the Legal and Regulatory Framework

The legal framework governing TDM, AI training, and data protection in the EU remains in active development. Implementing measures under the EU AI Act, evolving GDPR enforcement guidance from national data protection authorities and the European Data Protection Board, and emerging jurisprudence on the application of copyright law to AI training processes are each capable of materially affecting the conditions under which corpus construction is lawful. A standing process must be established within the project governance structure to monitor these developments and to propagate material changes — including revised interpretations of TDM exceptions, new opt-out mechanism requirements, and updated data-protection obligations — into corpus governance policies, dataset documentation, and the applicable legal basis assessments for subsequent training cycles. Legal and compliance review gates within the model lifecycle process must be configured to incorporate updated interpretations as authoritative guidance and case law mature.

10.2 Copyright and licensing checklist

The integration of scientific literature into large language model (LLM) training pipelines raises a distinct set of copyright and licensing obligations that must be addressed systematically prior to corpus construction and, thereafter, throughout the lifecycle of the resulting AI system. The following checklist operationalises the applicable legal framework — principally the EU Directive on Copyright in the Digital Single Market (CDSM Directive, 2019/790), the EU AI Act, and the Creative Commons licence suite — into a structured set of compliance requirements. Each element represents a necessary condition for legally sound corpus assembly and model deployment within the European regulatory context.

10.2.1 Source Inventory and Content Classification

Prior to the commencement of corpus construction, all planned data sources must be comprehensively catalogued. The inventory shall encompass open-access publisher platforms, institutional and subject-specific repositories (such as PubMed Central and Europe PMC), preprint servers, licensed subscription databases, and any collections assembled through automated web crawling. For each source, a content classification must be assigned that identifies the applicable rights regime — public domain, Creative Commons (CC) licence, subscription-controlled access, preprint deposit, or uncertain provenance — so as to enable differentiated legal treatment at each subsequent stage of the compliance workflow.

The technical modality through which content is accessed must also be documented as part of the inventory, since the access pathway — whether through a publisher API, a bulk dataset transfer, an institutional subscription arrangement, or a direct repository download — may independently affect the conditions under which text and data mining (TDM) exceptions are available. This documentation forms the evidentiary basis for all downstream legal assessments.

10.2.2 Confirmation of Lawful Access

Lawful access to the source content is a threshold requirement under the CDSM Directive. Before any act of reproduction for TDM or AI training purposes is initiated, the lawful accessibility of each source must be affirmatively confirmed. Lawful access may derive from a valid institutional

subscription, an open-access designation by the publisher or depositing author, or demonstrable public availability under terms that do not restrict such use. The absence of confirmed lawful access forecloses reliance on both Article 3 and Article 4 of the CDSM Directive, irrespective of the nature of the entity conducting the TDM.

10.2.3 Identification of the Applicable Legal Basis

For each source or discrete corpus segment, a specific and documented legal basis for reproduction must be established. The available bases under the EU framework are: the TDM exception for scientific research under Article 3 CDSM (applicable to qualifying research organisations and cultural heritage institutions); the general TDM exception under Article 4 CDSM (applicable to all natural and legal persons, subject to rights-reservation conditions); a Creative Commons licence of a specified variant; a bespoke contractual licence negotiated directly with the rightsholder; or established public-domain status of the content.

Where none of the foregoing bases applies — most critically, where no TDM exception is available and no open licence or bespoke agreement covers the intended use — reliance on alternative legal theories such as fair use must be assessed with care, given that EU law does not recognise a general fair use doctrine and that training on uncleared copyrighted material is treated as prima facie infringing under prevailing EU legal analysis. Any reliance on jurisdiction-specific exceptions beyond the CDSM framework must be separately substantiated and formally documented.

10.2.4 Verification of CDSM Article 3 and Article 4 Eligibility

Where reliance on a statutory TDM exception is intended, the specific eligibility conditions of the applicable article must be assessed per project and per corpus segment rather than assumed at an organisational level. Article 3 CDSM affords research organisations and cultural heritage institutions a mandatory exception for TDM conducted for the purposes of scientific research; this exception operates without a rightsholder opt-out mechanism and is subject to conditions of secure storage and non-commercial purpose. Commercial entities and mixed-purpose developers must instead assess eligibility under Article 4 CDSM, which permits TDM on lawfully accessible works provided that the rightsholder has not reserved their rights in a machine-readable or otherwise appropriate manner.

The assessment must account for the organisational status of the entity, the stated purpose of the training run (research, commercial, or hybrid), and the applicable storage and access conditions. Where a project involves collaboration between a qualifying research institution and a commercial partner, the eligibility of each entity with respect to each corpus segment must be evaluated independently, as the exception does not transfer by virtue of collaboration alone.

10.2.5 Detection and Enforcement of Machine-Readable Rights-Reservation Signals

Compliance with the Article 4 CDSM exception requires that any works for which the rightsholder has reserved rights in an appropriate manner be excluded from TDM pipelines. Automated detection mechanisms must therefore be implemented to identify rights-reservation signals before corpus ingestion, including publisher TDM opt-out declarations, robots.txt directives restricting automated access, and structured opt-out metadata embedded in content interfaces or access systems. Works carrying such signals must be excluded at the corpus-construction

stage; post-hoc exclusion following ingestion is insufficient for legal compliance and does not retroactively legitimise the initial reproduction.

10.2.6 Creative Commons Licence Variant Identification and Condition Compliance

Variant Identification

Each CC-licensed work included in the corpus must have its specific licence variant identified and recorded at source level, given that the legal implications for AI training differ materially across the available variants. The variants — CC BY, CC BY-SA, CC BY-NC, CC BY-NC-SA, CC BY-ND, and CC BY-NC-ND — impose distinct attribution, ShareAlike, NonCommercial, and NoDerivatives conditions, each of which operates differently in the context of corpus assembly and model training. Licence identification must occur at corpus-construction stage and be tied to source-level provenance metadata to enable downstream filtering and condition-specific compliance.

NoDerivatives (ND) Licensed Works

Works distributed under CC BY-ND or CC BY-NC-ND licences must be excluded from training corpora where the training process or the resulting model outputs may be characterised as derivative works within the meaning of applicable copyright law. In the absence of a separate, explicit licence permitting derivative use, obtained directly from the rightsholder, inclusion of ND-licensed content carries a material legal risk. It is acknowledged that the question of whether AI training constitutes the creation of a derivative work remains legally contested in EU jurisprudence; however, the precautionary principle supports exclusion of ND-licensed content absent a specific authorisation.

NonCommercial (NC) Licensed Works

Content governed by NonCommercial CC licence variants — CC BY-NC, CC BY-NC-SA, and CC BY-NC-ND — must be segregated into corpora designated exclusively for non-commercial training runs. Inclusion of NC-licensed content in datasets used to train commercially deployed AI products or services constitutes a breach of the NonCommercial condition and cannot be cured retroactively. Where the commercial or non-commercial character of the intended deployment is uncertain or hybrid in nature, NC-licensed content should not be relied upon unless an independent statutory basis independently permits the use.

Attribution and ShareAlike Conditions

CC BY and CC BY-SA licences impose attribution obligations that must be satisfied in any public dissemination of derived datasets, trained models, or outputs considered adaptations under applicable law. A mechanism for systematic attribution must be incorporated into corpus documentation and into any artefacts published in connection with the project. Additionally, the ShareAlike condition may require that publicly distributed models or datasets trained predominantly on SA-licensed content be released under an equivalent or compatible open licence. The applicability of ShareAlike conditions to trained model weights and output datasets is a legally unsettled question that must be assessed by qualified legal counsel prior to any public deployment.

10.2.7 Review of Publisher-Specific TDM and API Conditions

Each publisher's dedicated TDM policy and, where applicable, API usage terms must be independently reviewed to determine the conditions under which authorised users may conduct TDM under institutional subscription arrangements. Such policies commonly specify the access

mechanism required (for example, a dedicated API key or a separately concluded TDM contract), as well as rate limits, scope restrictions, and permitted purposes. Publisher TDM policies frequently impose conditions that are supplementary to, or more restrictive than, the CDSM statutory baseline; accordingly, both the statutory exception and the applicable contractual terms must be satisfied concurrently and their respective requirements documented.

10.2.8 Verification of Redistribution Restrictions on Full-Text Corpora

Publisher licences and institutional subscription agreements commonly prohibit or significantly restrict the redistribution of full-text corpora assembled from subscription content. These redistribution constraints must be identified during corpus intake, recorded in the dataset register, and actively enforced in any downstream data-sharing or publication workflows. It is to be noted that non-redistribution obligations apply independently of the purpose for which the corpus is subsequently used: the act of making the assembled corpus available to third parties — including project collaborators, subcontractors, or open-access repositories — may independently constitute a breach of the applicable licence, regardless of whether the corpus itself is used for training purposes.

10.2.9 Verification of Downstream Constraints on Trained Models and Outputs

Beyond the conditions governing corpus assembly, licence terms and publisher TDM policies may independently impose constraints on the permissible uses of AI models trained on the licensed content. Such downstream constraints may include restrictions on commercial deployment of the trained model, conditions on the sharing or publication of model weights, and limitations on the scope of use to purposes within the licensed perimeter. These constraints must be captured at corpus-intake stage as part of the provenance record and systematically communicated to product and deployment teams as binding operational parameters governing the permitted use of any AI systems derived from the licensed corpus.

10.2.10 Dataset Register and Provenance Documentation

A machine-readable dataset register must be maintained for the duration of the project and beyond, documenting — per source or corpus segment — the legal basis for use, the applicable CC variant or licence identifier, any rightsholder opt-out signals detected and applied, redistribution and reuse restrictions, the technical access modality, and the date and version of acquisition. This register constitutes the primary provenance record for the purposes of subsequent copyright audits, regulatory inspections, and the transparency obligations imposed by Article 53 of the EU AI Act on providers of general-purpose AI models. The register must be kept current as sources are added, removed, or reinterpreted in light of evolving legal guidance.

10.2.11 Alignment with Article 53 EU AI Act: Copyright-Compliance Policy and Training-Content Summaries

Providers of general-purpose AI (GPAI) models falling within the scope of the EU AI Act are required under Article 53 to implement and document a policy for compliance with EU copyright law, with particular regard to the rightsholder opt-out mechanism under Article 4 CDSM. This policy must specifically address scientific literature as a training source category and must

operationalise the detection, exclusion, and documentation of opted-out works across all data-ingestion pipelines.

Article 53(1)(d) further requires GPAI model providers to publish training-content summaries of sufficient detail to enable rightsholders to assess whether their works were included in the training corpus and whether opt-out and licensing conditions were duly respected. For corpora comprising scientific literature, such summaries should enumerate the major publishers and repositories accessed, the principal subject domains covered, the temporal scope of the content, and the content categories represented, while protecting any legitimately confidential trade-secret information. Summaries must be substantive and specific in their informational content; formal or generic disclosures that do not enable meaningful rightsholder verification do not satisfy the statutory transparency obligation.

10.2.12 Takedown and Opt-Out Handling Process for Future Model Versions

An operational process must be established and maintained enabling rightsholders to request the exclusion of their works from future training runs and, where technically and legally feasible, from future model versions or updates. This process must be formally documented, publicly accessible, and integrated into the dataset governance lifecycle to ensure that exclusion requests are logged, traceable to the specific works concerned, actioned within defined and reasonable timeframes, and reflected in the composition of subsequent training cycles. The existence and operability of the process is itself a component of the copyright-compliance policy required under Article 53 and should be described in the relevant transparency documentation.

10.3 Privacy, security, and governance checklist

The present checklist constitutes a structured compliance instrument addressing the principal privacy, security, and governance obligations that arise when scientific literature — including medical and clinical publications — is processed as training data for large language models (LLMs). Its scope encompasses the full data lifecycle from corpus assembly to model deployment, and it is intended to operate as an operational reference within the broader legal framework established by Regulation (EU) 2016/679 (GDPR), the EU AI Act (Regulation (EU) 2024/1689), and, where applicable, sector-specific health-data legislation. The checklist is organised thematically, with each subsection addressing a discrete compliance domain. Compliance with the requirements set out herein is to be treated as a prerequisite for, or a concurrent condition of, each phase of the training pipeline.

10.3.1 Corpus Mapping and Identification of Personal-Data-Bearing Segments

Prior to any ingestion of scientific literature into the training pipeline, the corpus must be systematically mapped to identify all components that embed personal data. Such components include, but are not limited to, individual case reports, qualitative study records, patient-level supplementary datasets, author biographical information, and acknowledgement sections. The mapping exercise must extend to both structured data fields (e.g., metadata schemas, supplementary tables) and unstructured free text, and must distinguish between corpus segments in which personal data appears incidentally and those in which it constitutes the primary data type. Segments identified as potentially personal-data-bearing must be isolated for targeted compliance review before ingestion commences. Failure to perform this mapping at the

outset creates downstream risk of unlawful processing that may be difficult to remediate once model training has begun.

10.3.2 Flagging of Special-Category Data

Health data, genetic data, and all other categories of personal data enumerated in Article 9 GDPR must be affirmatively flagged whenever they are encountered during corpus review. These categories arise with particular frequency in clinical journals, trial reports, and case-report sections of biomedical literature, and their presence triggers a heightened compliance pathway that is legally and procedurally distinct from the handling of ordinary personal data. Automatic screening tooling must be configured to detect terminology indicative of special-category data — including clinical, diagnostic, biometric, and psychometric content — and to route flagged segments to human review before they are cleared for inclusion in the training set. For each flagged segment, a documented legal basis under Article 9(2) must be established, together with the mandatory safeguards applicable to that basis. The default treatment for any ambiguous segment must be to classify it as special-category until the contrary is established.

10.3.3 Data Protection Impact Assessment

A Data Protection Impact Assessment (DPIA) pursuant to Article 35 GDPR is mandatory where the training corpus involves personal data or where the AI system under development is likely to produce outputs that may meaningfully affect individuals. The DPIA must be initiated at project inception rather than retroactively, and must be re-assessed following any material change to corpus scope, model architecture, or deployment context. The assessment must systematically describe the envisaged processing operations and their purposes, evaluate the necessity and proportionality of the processing in relation to its objectives, and assess the risks to the rights and freedoms of data subjects. DPIA findings are to inform go/no-go decisions at each stage gate prior to training and must be maintained as living documentation throughout the full model lifecycle. Where the DPIA identifies high residual risk that cannot be mitigated, the supervisory authority must be consulted in accordance with Article 36 GDPR before processing proceeds.

10.3.4 Determination of Controller and Processor Roles

For each training activity involving personal data, the participating organisation must formally determine whether it operates as a data controller — that is, an entity that determines the purposes and means of processing — or as a data processor acting on the documented instructions of another controller. Each role carries materially different obligations under the GDPR, and a conflation of the two creates regulatory risk. Where multiple entities jointly determine processing purposes and means — as may occur in a research consortium or a public–private partnership — a joint-controller arrangement must be established pursuant to Article 26 GDPR, and the respective responsibilities of each joint controller must be defined in a transparent internal arrangement and communicated to data subjects. Role determinations must be documented and revisited whenever the organisational, contractual, or technical configuration of the project changes in a manner that may affect the allocation of processing responsibilities.

10.3.5 Record of Processing Activities

All AI training activities involving personal data must be recorded in the organisation's Record of Processing Activities (RoPA) maintained pursuant to Article 30 GDPR. Each RoPA entry for a training activity must capture, at minimum, the categories of personal data processed, the categories of data subjects affected, the purposes of the processing, the applicable retention periods, and the recipients or categories of recipients to whom data may be disclosed. RoPA

entries must cross-reference the associated DPIAs, lawful-basis determinations, and any data processing agreements concluded with third-party processors. The RoPA must be maintained in a current state and made available to competent supervisory authorities upon request without undue delay.

10.3.6 Lawful Basis under Article 6 GDPR

A specific and documented lawful basis under Article 6(1) GDPR must be identified for each distinct processing activity within the training pipeline; where different stages of the pipeline rely on different bases, each must be recorded separately. For scientific research training activities, the most likely applicable bases are legitimate interests under Article 6(1)(f) or the scientific research ground available under national implementing legislation. Where legitimate interests are invoked, a balancing test or Legitimate Interest Assessment (LIA) must be conducted and documented, weighing the organisation's interests against the rights and reasonable expectations of the data subjects. Where consent is relied upon, its scope, specificity, and revocability must be assessed for compatibility with iterative and incremental training workflows, which may render retroactive withdrawal technically complex. Any reliance on consent as the operative basis for a continuous training pipeline should therefore be evaluated with particular care.

10.3.7 Legal Condition for Special-Category Data under Article 9(2) GDPR

Where special-category data is present in the training corpus, a specific exception under Article 9(2) GDPR must be identified and documented for each processing activity. The most commonly applicable conditions in a scientific research context are explicit consent under Article 9(2)(a) or processing for scientific research purposes with appropriate safeguards pursuant to Article 9(2)(j), as implemented by the relevant Union or Member State law. The chosen Article 9(2) condition must be documented alongside the corresponding Article 6 basis, and the supporting legal rationale must be recorded in both the DPIA and the RoPA. Reliance on the scientific research exception under Article 9(2)(j) is conditional upon demonstrable alignment with applicable Union or national law and upon the application of the additional safeguards required by Article 89 GDPR. Organisations must not assume that the scientific character of source material in and of itself satisfies the conditions of this exception; a case-by-case legal assessment is required.

10.3.8 Application of Article 89 Safeguards

Processing of personal data for scientific research purposes under Article 89 GDPR requires the implementation of technical and organisational safeguards that are proportionate to the re-identification risk presented by the specific corpus. Required safeguards include pseudonymisation, access restrictions, and data minimisation at the corpus level prior to ingestion, as well as access controls and audit logging at the training infrastructure level, and memorisation-mitigation measures at the model level. Article 89 safeguards must be operationalised at each of these three levels and must not be limited to pre-processing steps alone. Derogations from data-subject rights that are available under Article 89 — such as the suspension of the right of access or the right to erasure in certain research contexts — must be individually justified and formally documented; they do not, however, eliminate transparency obligations, which must be satisfied through appropriate privacy notices and information disclosures to data subjects.

10.3.9 Data Minimisation and Purpose Limitation

Personal data incorporated into the training corpus must be limited to what is strictly necessary for the defined training objective, in accordance with the data minimisation principle under Article 5(1)(c) GDPR. Deliberate selection and cleaning procedures must be employed to exclude personal identifiers that are superfluous to the training purpose, and these procedures must be documented in the dataset's provenance record. Where personal or patient data originally collected for a different purpose — such as clinical care or a prior research study — is to be reused for AI training, a purpose-compatibility assessment under Article 6(4) GDPR must be conducted and documented before reuse commences. This assessment must evaluate the link between the original and the new purpose, the context in which the data was collected, the nature of the data, the possible consequences of the intended further processing, and the existence of appropriate safeguards. Purpose-limitation controls must also be enforced technically, ensuring that personal-data-bearing corpus segments are not repurposed for processing activities beyond the scope for which they were approved.

10.3.10 Preference for Anonymised or Aggregate Data

In accordance with the principle of data protection by design and by default under Article 25 GDPR, anonymised or aggregate alternatives to identifiable personal data must be preferred wherever the training objective can be achieved without recourse to personal data. The decision to use personal data in lieu of anonymised alternatives must be affirmatively justified and recorded, demonstrating that anonymisation was considered and found to be technically insufficient for the specific use case. Organisations are accordingly required to evaluate privacy-preserving architectures — including synthetic data generation, federated learning, and differential privacy — as primary options before centralising raw personal data in a training environment. The output of this evaluation must form part of the DPIA and the dataset documentation. Where anonymisation is technically feasible but operationally inconvenient, convenience is not a legally sufficient basis for the preferential use of identifiable data.

10.3.11 Re-identification Risk Assessment

Published de-identification of scientific literature cannot be assumed to satisfy the legal standard for anonymisation under the GDPR, which requires that re-identification be reasonably impossible given all means likely to be used. Residual re-identification risk must therefore be independently assessed, with particular attention to scenarios in which multiple corpus sources are combined or linked to external datasets, which can materially increase the probability of re-identification through inference or linkage attacks. The risk assessment must account for advances in inference techniques and for the information density of the specific corpus — rare-disease case reports, for example, present substantially higher re-identification risk than general population studies. The findings of the re-identification risk assessment must feed directly into DPIA conclusions, anonymisation standards, and access-control decisions, and the assessment must be repeated if the corpus composition, linking environment, or state of the art in re-identification techniques changes materially.

10.3.12 Confidentiality and Access-Control Measures

Access to personal-data-bearing corpus segments must be restricted to authorised personnel through role-based access controls. Access logs must be maintained for audit purposes and must record the identity of the accessor, the time and nature of the access, and the corpus segment accessed. Technical barriers — including encrypted storage, segmented processing environments, and network-level access restrictions — must prevent unauthorised access to raw

corpora and training artefacts that contain identifiable data. All staff and third parties with access to personal-data-bearing training materials must be subject to documented confidentiality obligations that are contractually enforceable. These obligations must be proportionate to the sensitivity of the data involved and must remain in force for as long as the relevant training artefacts are retained.

10.3.13 Pseudonymisation and Anonymisation Controls

Privacy-enhancing technologies must be incorporated into the training pipeline to reduce the risk of memorisation and to prevent the leakage of sensitive personal data through model outputs. Applicable techniques include pseudonymisation, anonymisation, differential privacy mechanisms applied during training, and output-level safe-completion filters. Any anonymisation applied must meet the technical and legal standards applicable in the relevant jurisdiction — including, where applicable, the HIPAA Safe Harbor standard for US-origin health data — and residual re-identification risk must be assessed after application to verify that the anonymisation threshold has been reached. Where pseudonymisation is used in lieu of full anonymisation, it must be accompanied by secure key management protocols and access restrictions on re-identification keys, ensuring that the combination of pseudonymised data and key material remains unavailable to unauthorised parties. The use of pseudonymised data does not reduce the applicable GDPR obligations, which apply in full where re-identification remains reasonably possible.

10.3.14 Retention Limits for Raw Corpora and Training Artefacts

Documented retention policies must be established for all data holdings created or maintained in connection with the training pipeline, including raw corpora, intermediate caches, embeddings, checkpoints, and other training artefacts that may contain identifiable personal data. Each retention policy must specify the maximum retention period applicable to each category of holding, the trigger events for deletion or de-identification, and the responsible role for ensuring that deletion is effected. The duration for which identifiable data remains accessible must be minimised to the period strictly necessary for the legitimate training purpose; archival beyond active training must be separately justified and subject to access controls equivalent to those applicable during active use. Retention schedules must be aligned with any sectoral health-data retention requirements applicable to the source data, and must be consistently reflected in the RoPA. Retention policies must be reviewed and updated at each project phase transition.

10.3.15 Data-Subject Rights Handling Capability

Operational processes and tooling must be in place to receive and respond to data-subject requests across all components of the training environment, including raw corpora, access logs, and derived artefacts. The applicable rights — including the right of access under Article 15, the right to erasure under Article 17, the right to rectification under Article 16, and the right to object under Article 21 — must each be supported by documented handling procedures that specify response timelines, responsible personnel, and escalation paths. Organisations must assess, and document in their privacy notices, the technical feasibility and limitations of locating and removing a specific individual's data from training artefacts, given that neural model weights do not permit deterministic data extraction. Erasure requests must trigger a documented workflow covering corpus removal, artefact updates, and, where technically feasible, assessment of whether the affected model version must be deprecated or retrained. Where technical limitations preclude complete erasure from trained model weights, these limitations must be disclosed to data subjects and to the supervisory authority where required.

10.3.16 Ethics and Confidentiality Approvals for Secondary Use of Patient-Level Data

Secondary use of patient-level data for AI training requires prior verification of applicable ethics-committee approvals, confidentiality-advisory-group authorisations, or equivalent regulatory research exemptions. This requirement applies regardless of whether the patient data has been de-identified, as the regulatory status of secondary use is determined at the point of data access and not solely by the form in which the data is held. Where existing approvals do not explicitly cover AI training as a use case — which is frequently the case for approvals granted before the current generation of LLM training methodologies — additional ethical review must be sought before training commences; reliance on an implicit extension of an existing approval scope is not legally sufficient. The scope and conditions of any approval obtained, together with any restrictions on data sharing or downstream model deployment, must be recorded in the project governance file and enforced in product design decisions.

10.3.17 Vendor Contractual Controls for External Processing

Any external processing of personal or health data in connection with AI training must be governed by a Data Processing Agreement (DPA) satisfying the requirements of Article 28 GDPR, or by a Business Associate Agreement (BAA) where US health-data law applies. The DPA or BAA must specify, at minimum, the permitted processing purposes, sub-processor restrictions, audit and inspection rights, data-return or deletion obligations upon termination, and breach-notification timelines. Subject-matter, duration, nature, and purpose of processing, as well as the type of personal data and categories of data subjects, must be expressly defined in the agreement in conformity with Article 28(3) GDPR. Prior to engagement, documented due diligence must be conducted on the vendor's technical and organisational security measures, and this due diligence must be refreshed periodically or upon any material change to the vendor's processing environment. Sub-processors may be engaged only where the primary processor has obtained prior specific or general written authorisation from the controller in accordance with Article 28(2).

10.3.18 Governance-Body Review and Go/No-Go Decision

The corpus inventory, associated copyright analysis, DPIA findings, and ethics-approval status must be presented to the designated AI governance body or steering committee for a formal go/no-go decision before training commences. The governance body must be constituted with sufficient cross-functional representation — spanning legal, privacy, research, and AI compliance functions — to assess the legal, ethical, and technical dimensions of the proposed training run in an integrated manner. A unidimensional review conducted solely by technical or legal personnel, without the participation of the other functions, does not satisfy this requirement. Go/no-go decisions must be formally recorded with a statement of rationale, any dissenting views, and the conditions under which the decision is subject to reassessment. Conditional approvals must specify the actions required and the timeline within which compliance must be demonstrated before training proceeds.

10.3.19 Recording of Approvals, Constraints, and Monitoring Requirements

All approvals granted by the governance body, together with any applicable constraints — such as restrictions to non-commercial use, limitations on deployment contexts, or time-limited data access authorisations — must be formally documented in the system's technical file and, where applicable, in the AI Act conformity assessment record. Monitoring requirements specified at the approval stage — including red-teaming schedules, memorisation-auditing frequency, data-subject rights review cycles, and

incident-response timelines — must each be assigned to a named responsible role, ensuring clear accountability. The approval record must be version-controlled so that subsequent modifications to constraints or monitoring requirements are traceable. Where monitoring findings indicate a material change in risk or a failure to meet an approval condition, a reassessment by the governance body must be triggered without undue delay.

10.3.20 Auditability of Filtering, Transformations, and Approvals

All filtering and transformation steps applied to the training corpus — including rights-reservation exclusions, personal-data removal, deduplication, sampling, and any other modification of the raw source material — must be logged in the dataset's provenance record in a manner that enables retrospective audit and regulatory inspection. Approval decisions, DPIA outcomes, and governance-body records must be retained in a durable, structured form that supports third-party audit throughout the full model lifecycle and not merely during the active training period. Provenance records must be machine-readable and structured to facilitate automated compliance checks, including cross-referencing against opted-out source material, copyright catalogues, and applicable exclusion lists. The completeness and integrity of provenance records must be verified at regular intervals and upon any change to the dataset or training pipeline. Organisations should treat provenance documentation as a first-class compliance artefact, equivalent in importance to the technical file required under the EU AI Act for high-risk AI systems.

10.4 Pre-training or pre-sharing checklist

This checklist constitutes an operative compliance instrument designed to be executed prior to the initiation of any training run involving medical scientific literature or prior to the sharing of a configured corpus with partner organisations. It synthesises the legal, technical, and governance obligations identified throughout the preceding sections of this deliverable and structures them into a sequential, auditable procedure. Compliance with each item is a prerequisite for a formally authorised go/no-go decision; no item may be deferred post-hoc without documented justification and governance approval.

10.4.1 Scope Definition and Risk Classification

Definition of System Type and Primary Use Case

Prior to corpus selection or any data-acquisition activity, the system under development must be formally characterised with respect to its architectural type and operational purpose. This characterisation determines the entirety of the applicable regulatory regime. A general-purpose AI (GPAI) model, a domain-specific language model, and a retrieval-augmented generation (RAG) architecture each attract distinct obligations under Regulation (EU) 2024/1689 (the EU AI Act), under the Directive on Copyright in the Digital Single Market (CDSM Directive), and under the GDPR. Accordingly, the system type must be specified in the technical documentation before any other step in this checklist is undertaken.

In parallel, all primary deployment scenarios must be stated explicitly and with sufficient specificity. Intended use cases such as automated summarisation of biomedical literature, clinical question-answering support, or research-assistance generation are not interchangeable from a regulatory standpoint: each carries different risk implications and may engage different

normative instruments. Use-case specificity is not a formality but a material determinant of applicable obligations, including those concerning data quality, model transparency, and conformity assessment under the EU AI Act.

Classification Against the EU AI Act Risk Taxonomy

Each prospective system must be mapped to the risk tier established by the EU AI Act prior to training. Systems deployed in clinical decision support, patient triage, employment screening, or educational assessment fall within Annex III and are classified as high-risk, thereby triggering enhanced obligations with respect to data-quality management, technical documentation, human oversight, conformity assessment, and post-market monitoring. This classification is not a static determination: it must be revisited whenever the deployment context, intended user base, or functional scope of the system undergoes substantive change. Where reclassification is required, the full set of enhanced obligations attaches from the point of reclassification, and prior documentation must be updated accordingly.

Determination of Research or Commercial Deployment Context

A clear determination as to whether the project constitutes genuine scientific research — and thereby potentially qualifies for the text and data mining (TDM) exception under CDSM Article 3 and the scientific research derogation under GDPR Article 89 — or constitutes a commercial deployment must be made and documented before any data-acquisition or training activity commences. This determination governs which legal bases are available for both copyright reliance and personal data processing. Reliance on research exemptions for a system that is subsequently commercialised constitutes a change of purpose that prospectively invalidates the original legal basis; any such transition must therefore be identified in advance and documented with an assessment of the consequential legal obligations that arise.

10.4.2 Corpus Inventory and Provenance Mapping

Enumeration of Datasets and Source Types

A comprehensive inventory of all planned corpus sources must be produced before training commences. This inventory must cover the full range of source categories under consideration, including open-access repositories, commercial publisher platforms, preprint servers, institutional repositories, and any materials acquired through automated web retrieval. Each source category must be separately classified by content type and by the access pathway through which the materials were obtained, given that TDM permissions and licence conditions frequently differ per channel even where the same underlying publication is available through multiple routes.

Establishment of a Machine-Readable Provenance Register

A machine-readable dataset register must be established before training commences, capturing, at a per-source level of granularity, the acquisition method, date of retrieval, applicable licence or legal basis for TDM, and any rights-reservation signals detected at the time of ingestion. Where corpus subsets contain personal or sensitive data — including, without limitation, clinical case reports, qualitative social-science studies incorporating identifiable subjects, or acknowledgement sections naming individuals — those subsets must be tagged for differentiated processing and subjected to heightened oversight throughout the pipeline.

Provenance records must be treated as living artefacts rather than static outputs. They must be retained throughout the system lifecycle and made available on demand to support audit processes, the fulfilment of opt-out requests, and transparency reporting under EU AI Act Article

53. Failure to maintain current provenance records constitutes a compliance gap that cannot be retrospectively remediated by reference to snapshots produced at an earlier point in the lifecycle.

10.4.3 Copyright and License audit

Classification of Corpus Segments by Licence Type

Automated detection of licence variants and publisher rights-reservation signals must be applied to the full corpus to classify each segment by its applicable intellectual property status before any training run is initiated. Content released under Creative Commons Attribution (CC BY) or Attribution-ShareAlike (CC BY-SA) licences is, in principle, suitable for training purposes. Content released under Attribution-NonCommercial (CC BY-NC) or Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) licences is restricted to demonstrably non-commercial uses, and the project's deployment context — as established under item A.3 — must be verified against that restriction. Content released under Attribution-NoDerivatives (CC BY-ND) or Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) licences must be excluded from corpora intended for training where model outputs may constitute derivative works within the meaning of applicable copyright law.

Exclusion of Opted-Out Works Under CDSM Article 4

For corpus elements relied upon under the CDSM Article 4 commercial TDM exception, automated exclusion of works in respect of which rightsholders have issued machine-readable opt-outs is mandatory and not discretionary. Such opt-outs may be expressed through robots.txt directives, publisher metadata flags, or dedicated TDM policy portals operated by rights-holder associations. The exclusion pipeline must be capable of processing each of these signal types and must be executed against the current state of rights-reservation signals immediately prior to ingestion, as opt-out status may change between the date of corpus scoping and the date of training.

Copyright Compliance Policy for GPAI Providers

Organisations developing GPAI models must verify that a copyright compliance policy specifically addressing scientific-literature sources, as required by EU AI Act Article 53(1)(c), is operative and documented before training commences. The existence of such a policy at an abstract level is insufficient; its application to the specific corpus under consideration must be demonstrable and must be referenced in the pre-training governance submission.

10.4.4 Data protection and ethics Assessment

Data Protection Impact Assessment

A Data Protection Impact Assessment (DPIA) is required prior to training whenever the corpus contains personal data — whether that data constitutes the primary purpose of the corpus or is incidentally present — or where model outputs could meaningfully affect identifiable individuals. For corpora incorporating patient-level data or sensitive health information, a formal review by an ethics committee or confidentiality advisory body, conducted in accordance with the applicable national health-data framework, must be completed and documented before any data ingestion occurs. The DPIA must not be treated as a post-training formality; its function is to inform corpus design and processing decisions before irreversible training commitments are made.

Determination of Lawful Basis Under GDPR

The lawful basis for processing under GDPR Article 6 must be determined and recorded for each personal data category present in the corpus. Where the corpus includes special-category data as defined in GDPR Article 9 — including health data, biometric data, or data revealing racial or ethnic origin — a specific Article 9(2) condition must additionally be identified and documented. Where processing is conducted on the basis of scientific research grounds under Articles 6(1)(e) or 9(2)(j), the technical and organizational safeguards mandated by Article 89, including pseudonymization, access controls, and purpose limitation, must be implemented and documented prior to training.

10.4.5 Technical Filtering and Transformation

Data Minimisation Through Active Removal or Masking

Data minimisation requires the active removal or masking of identifiable personal data from training corpora wherever such data is not strictly necessary for the intended model capability. The default position must be exclusion unless a documented justification supports the retention of identifiable data elements. Published de-identification measures applied to source materials cannot be assumed sufficient: residual re-identification risk must be independently assessed, in particular where corpora are combined with external datasets that may enable linkage attacks not foreseeable from analysis of any single source in isolation.

Exclusion Pipeline for Non-Compliant Sources

A programmatic exclusion pipeline must be implemented and verified prior to any training run. This pipeline must remove, at a minimum: works for which rightsholders have issued valid opt-outs under any applicable instrument; content licensed under NoDerivatives terms where no alternative legal basis is available; and materials obtained through automated retrieval that violated platform terms of service at the time of acquisition. The pipeline must be documented and reproducible, such that its outputs can be reconstructed for audit purposes at any subsequent point in the system lifecycle.

Systematic Deduplication

Systematic deduplication of corpus documents is required to mitigate verbatim memorisation of individual works and to prevent the statistical over-representation of specific publications, which can distort model behaviour and amplify the risk of copyright-infringing regurgitation in model outputs. Deduplication must be applied at the document level as a baseline measure. Supplementary controls, such as maximum token-span limits per source document or stratified sampling schemes that enforce representation ceilings across publishers or journals, may be applied to further reduce verbatim retention risk.

Logging of Filtering and Transformation Operations

Every filtering, exclusion, deduplication, and transformation operation applied to the corpus must be logged with sufficient granularity to permit full reconstruction of the final training dataset's composition at any subsequent point. Log records must be retained and made accessible to support audit requirements under both the EU AI Act and GDPR Article 30, which mandates the maintenance of records of processing activities. Informal or incomplete logging is insufficient; the logging regime must be implemented as a systematic and verifiable process integrated into the data pipeline.

10.4.6 Governance Approvals

Formal Go/No-Go Decision by the AI Governance Body

The complete corpus inventory, copyright audit findings, DPIA conclusions, and ethics approvals must be submitted to the designated AI governance body or steering committee for a formal go/no-go decision before any training run is initiated or before any corpus is shared with partner organizations. This submission must be treated as a binding prerequisite: training runs initiated without a recorded governance approval are non-compliant irrespective of the substantive quality of the underlying documentation. Approval decisions must specify any constraints on permitted uses — for example, restrictions to non-commercial deployment or to defined deployment environments — and these constraints must be reflected in the system's technical documentation.

Alignment with Internal AI Policy and Sectoral Ethics Codes

Pre-training validation must confirm alignment with the organization's internal AI governance policy, including its risk-based framework for EU AI Act compliance, and with any applicable sectoral ethics codes, such as medical research ethics frameworks or clinical AI deployment guidelines. Where human-rights impact assessments are mandated for high-risk applications under applicable national or supranational instruments, the outcomes of those assessments must be integrated into governance checkpoint documentation and must be capable of being evidenced to supervisory authorities upon request.

10.4.7 Pilot Validation and Red-Teaming

Pilot-Run Validation on a Constrained Corpus Subset

Initial training must be conducted on a representative but constrained subset of the full corpus to evaluate baseline model behavior, data-quality characteristics, and pipeline integrity before committing to a full-scale training run. The pilot corpus must be sufficiently representative to surface systematic deficiencies in the data pipeline, in the license-classification logic, and in the filtering regime; it must not be designed to optimize pilot results at the expense of representativeness.

Copyright and Privacy Red-Teaming

Dedicated red-teaming exercises must be conducted against the trained model to probe for verbatim reproduction of scientific article passages and for the extraction of sensitive personal data. Red-teaming scope must encompass both copyright-regurgitation attack surfaces — that is, the ability to elicit substantial reproduction of specific works through targeted prompting — and privacy-leakage attack surfaces, involving the ability to extract individually identifiable information from the model's learned representations. Methodologies applied in academic medical-centre AI deployments, which involve systematic adversarial prompting designed to elicit memorised clinical text, provide an appropriate reference framework for the design of these exercises.

Remediation Following Red-Teaming Findings

Where red-teaming or pilot evaluation reveals levels of copyright regurgitation or personal data leakage that exceed acceptable thresholds, remediation must be implemented before proceeding to full-scale training or deployment. Remediation options include targeted corpus modification, the application of additional filtering layers, revised deduplication parameters, or the implementation of model-side safe-completion controls designed to interrupt outputs that match prohibited patterns. The choice of remediation method must be documented, and a

further validation cycle must confirm that identified deficiencies have been resolved to an acceptable level before the go/no-go decision for full training is revisited.

10.4.8 Transparency and Living Documentation

Preparation of a Training-Data Summary for GPAI Models

For GPAI models subject to EU AI Act Article 53(1)(d), a sufficiently detailed public-facing training-data summary must be prepared prior to model release. This summary must specify the major source categories, repositories, temporal ranges, and content types represented in the training corpus, calibrated to enable rightsholders to assess with reasonable confidence whether their works were used and whether applicable opt-outs and licence conditions were respected. The summary must protect legitimately confidential commercial information whilst providing a level of detail that satisfies the transparency objective of Article 53; a generic description of source categories without temporal or repository specificity does not meet this standard.

Living Documentation and Ongoing Regulatory Monitoring

Internal technical documentation — encompassing the dataset register, copyright analysis, DPIA, risk assessments, and governance approvals — must be treated as a living record and updated to reflect any changes in corpus composition, applicable legal interpretation, or regulatory guidance before each subsequent training cycle and before sharing with partner organisations. A monitoring function must be established with formal responsibility for tracking legal and regulatory developments relevant to AI training on scientific literature, including new supervisory guidance on copyright and AI, evolving GDPR enforcement positions, and updates to EU AI Act implementing acts or codes of practice. Material changes identified through this monitoring function must be fed into documentation and compliance processes on a timely and systematic basis.

11 Practical scenarios and decision rules

The legal framework outlined in the preceding sections must ultimately be applied to concrete and recurrent source-use situations that arise during corpus construction and handling. In practice, many compliance decisions are not taken in the abstract, but in relation to specific categories of source material that present recurring patterns of legal and operational complexity. The purpose of the present section is therefore to translate the general framework into scenario-based decision rules that support consistent and well-documented source assessment. The next table provides a comparative overview of the main scenarios considered in this deliverable and identifies, for each of them, the typical source context, the principal legal issue, the default compliance position, and the conditions that must be satisfied for use within the project.

Table 12. Scenario-based decision matrix for recurrent source-use cases.

Scenario	Typical source	Main legal issue	Default status	Conditions for use
Open-access article scenario	Openly accessible journal article or repository copy, potentially under a Creative Commons licence	Whether the applicable licence permits the intended training or reuse activity; whether attribution or other licence conditions are relevant	Generally permitted, subject to licence compatibility and documentation	Lawful access must be confirmed; licence terms must be checked; any relevant conditions must be documented
Licensed or subscription-based article scenario	Publisher-hosted full-text article or database-accessed publication available through subscription	Contractual restrictions, lawful access limits, TDM exception applicability, and possible opt-out or licence barriers	Restricted or conditional	Access route must be lawful; licence and subscription terms must be reviewed; use must fall within a valid exception or be separately authorised
Abstracts, metadata, tables, and figures scenario	Abstracts, bibliographic metadata, tables, charts, figures, or supplementary items extracted from publications or databases	Different content elements may attract different forms or levels of protection; reuse conditions may differ from those applicable to full text	Conditional	The specific content element must be identified; rights and restrictions must be assessed at element level; database or platform constraints must also be considered
Consortium-shared material scenario	Material circulated internally among consortium partners, including derived datasets, excerpts, or structured collections	Internal sharing does not eliminate upstream rights, privacy, or licensing constraints; onward use may exceed the original permitted scope	Conditional or restricted	Source origin and upstream rights must remain traceable; partner sharing must be consistent with the original legal basis and any applicable restrictions

11.1 Open-access article scenario

Open-access (OA) designation constitutes a materially significant, yet non-exclusive, legal basis for the incorporation of scientific literature into large language model (LLM) training corpora. The existence of an open-access licence establishes that the rightsholder has consented to a degree of downstream reuse, and this consent is relevant to the lawfulness of training use. However, OA status alone does not discharge all applicable compliance obligations. Organisations seeking to rely on openly licensed medical and scientific literature must

additionally establish a documented copyright position, conduct a licence-specific analysis, and demonstrate alignment with AI-specific regulatory requirements — most notably those arising under EU law. The assumption that open access is equivalent to unrestricted training use represents a material legal error that must be explicitly avoided in corpus governance frameworks.

11.1.1 Verification of Creative Commons Licence Variant

Not all open-access content is governed by identical licensing terms. Creative Commons licences operate as a family of legally distinct instruments, each imposing different conditions on downstream use. The principal variants in academic publishing — CC BY, CC BY-SA, CC BY-NC, and CC BY-ND — differ materially in the permissions they confer and the obligations they impose. The CC BY licence permits broad reuse, including for commercial purposes and in derivative works, subject only to attribution. CC BY-SA additionally imposes a ShareAlike condition, requiring that derivative works or adaptations be distributed under equivalent or compatible terms. The CC BY-NC licence restricts use to non-commercial contexts, a limitation with direct consequences for AI developers whose commercial deployment of a trained model may trigger the restriction. The CC BY-ND licence prohibits the creation of derivative works, which raises a fundamental question as to whether a trained model or a derived dataset constitutes a sufficiently transformed artefact to fall within this restriction.

Each OA source in the training corpus must therefore be evaluated individually against its precise licence variant, and a classification record maintained accordingly. Where NC or ND restrictions are present, organisations must either exclude the affected content from training pipelines or identify and establish an independent statutory exception — such as the text and data mining (TDM) exception under Article 4 of Directive (EU) 2019/790 — that would independently authorise the use irrespective of licence conditions. Reliance on a statutory exception in this context must itself be documented and its applicability assessed on the specific facts of each source.

11.1.2 Authorised Access Channels and Technical Compliance

The legal significance of OA status does not extend unconditionally to all methods by which content is obtained. Acquisition of open-access corpora must proceed through authorised access channels, including official bulk-download mechanisms, repository dataset dumps, or publisher-provided application programming interfaces (APIs). Ad hoc web scraping of OA repositories — even where the underlying content is freely licensed — may fail to satisfy the conditions for lawful access, particularly where the operator's terms of service restrict automated access or where crawl instructions have been disregarded. This distinction carries direct legal consequence: under EU law, the availability of a TDM exception and the right to invoke a Creative Commons licence both presuppose that the content has been accessed lawfully. An organisation that acquires OA content through means inconsistent with the applicable access terms cannot rely on those licences or exceptions as though its access had been authorised.

Technical access must, in all cases, comply with operator-imposed rate limits, crawl policies, and any machine-readable access directives — including, but not limited to, those expressed through robots.txt files. Non-compliance with such directives may independently constitute a breach of the terms governing lawful access, and may thereby defeat the preconditions necessary to invoke statutory TDM exceptions under EU law. Corpus engineering teams must therefore implement technical controls that enforce compliance with access instructions at the point of

acquisition, and must document the access method used for each source in the provenance record.

11.1.3 Attribution and ShareAlike Obligations

CC BY and CC BY-SA licences impose mandatory attribution requirements on any entity that exercises the permissions granted thereunder. In the context of training data use, this obligation must be operationalised at the corpus-engineering stage rather than addressed retrospectively. Where derived datasets or trained models are publicly shared or distributed — including, for example, through the release of a fine-tuned model accompanied by dataset documentation or a model card — attribution mechanisms must be in place that identify the source material with sufficient specificity. Acceptable attribution mechanisms include source listings in dataset documentation, annotated data cards, or structured metadata fields associated with model releases. Attribution retrofitted after the fact, without adequate provenance records, is unlikely to satisfy the attribution requirements imposed by CC BY or CC BY-SA instruments.

The ShareAlike condition warrants particular attention in the context of corpus composition. Where training corpora incorporate CC BY-SA material and the downstream artefacts produced — whether curated datasets or fine-tuned models — are publicly distributed and qualify as adaptations within the meaning of the applicable licence, the SA condition may require those artefacts to be released under equivalent or compatible licence terms. The legal characterisation of a trained model as an adaptation of its training data remains a contested and evolving question; however, organisations cannot assume a favourable outcome and must undertake a considered assessment at design stage. Where the intended deployment and sharing model is incompatible with SA obligations, the organisation must either adjust the composition of the training corpus to exclude CC BY-SA material, obtain a separate agreement with the relevant rightsholders, or restructure the licensing strategy for the derived outputs.

11.1.4 Provenance Tracking and Audit Capability

Open licensing does not diminish or displace the obligation to maintain systematic provenance records for each source incorporated into the training corpus. For every OA source, the provenance record must capture, at minimum: the access pathway and channel used for acquisition, the date of acquisition, the precise Creative Commons licence variant applicable at the time of access, and any opt-out, exclusion, or rights reservation indicators associated with the source or its publisher. Provenance documentation serves multiple operational functions: it underpins post-training audit capability; it enables rightsholder transparency in the event of a compliance enquiry; and it provides the foundation for iterative compliance management as sources change, licences are revised, or legal interpretations of applicable law evolve. The maintenance of provenance records is not, therefore, a purely administrative function — it is a prerequisite for the sustained legal defensibility of the training corpus.

11.1.5 AI Act Transparency Obligations

The open-access character of a training source does not exempt it from the transparency obligations introduced by the EU AI Act (Regulation (EU) 2024/1689). Article 53(1)(d) requires providers of general-purpose AI models to publish sufficiently detailed summaries of the content used for training; this obligation applies irrespective of whether the sources in question are openly licensed or acquired through statutory exceptions. OA corpora must accordingly be incorporated into training-data summaries and categorised by reference to the major repositories from which they were sourced, the time ranges covered, and the principal content types represented. Such summaries must be designed to enable rightsholders to verify

compliance with licence terms and applicable law, while remaining within the boundaries of legitimate trade secret protection where this is claimed. The AI Act's transparency requirements thus operate as an independent and concurrent obligation that must be addressed in parallel with licence compliance and TDM exception governance, and cannot be satisfied through generic or aggregated disclosures that obscure the provenance of material components of the training corpus.

11.2 Licensed or subscription-based article scenario

The present scenario applies where the scientific literature corpus to be used for LLM training or fine-tuning includes articles obtained through institutional subscriptions or publisher licences — arrangements that are standard in academic and research settings but that carry a distinct and often underestimated set of legal constraints when the intended downstream use is AI model development. The central analytical challenge of this scenario is to distinguish between the rights conferred by lawful access — i.e. the permission to read, download, and cite articles for conventional research purposes — and the materially different rights that would be required to reproduce, extract, and process those same articles at scale for use as training data. These two categories of right are legally separate, and the former does not, as a matter of law or contract, imply the latter.

11.2.1 Institutional Access and the Limits of Lawful Permission

An active subscription or institutional licence agreement confers upon the subscribing organisation — and upon the researchers designated as authorised users under that agreement — the right to access, read, and download the covered content for purposes consistent with the scope of the licence. This right of access is, however, strictly delimited by the terms of the agreement and does not, absent express contractual provision, extend to bulk reproduction, automated programmatic extraction, or the use of content as input material for the training of artificial intelligence or large language model systems. The scope of rights actually granted by any given subscription must therefore be independently assessed against the specific activities constituting the intended TDM or AI-training workflow, rather than assumed to encompass such activities by default.

A further consideration of legal significance is the effect of silence in the licence on TDM and AI-related uses. Where a licence agreement does not address text and data mining, machine learning, or AI training, that silence should be treated as a restriction rather than as implied permission. This interpretive posture reflects both the general principle that licences are construed narrowly against the licensee in matters of doubt and the specific policy framework established by the CDSM Directive, which sets out structured exceptions to exclusive rights rather than leaving the scope of permitted acts to implication. Accordingly, lawful access — whilst a necessary precondition for reliance on TDM exceptions under Article 4 of the CDSM Directive — is not in itself sufficient to authorise corpus ingestion for AI training purposes; additional contractual or statutory authorisation must be independently established.

11.2.2 Review of Publisher TDM Policies

Each major academic and scientific publisher maintains its own text and data mining policy, which may independently grant, conditionally permit, or expressly restrict TDM by authorised users, notwithstanding the rights conferred by the underlying subscription agreement. These policies vary substantially across publishers in both scope and formulation: some grant broad TDM permissions for non-commercial research purposes without further formality, whilst others impose detailed procedural requirements, technical channel constraints, or prior notification obligations. Because there is no harmonised publisher-side framework governing TDM permissions, each policy must be individually identified and reviewed before any corpus extraction activity is initiated for the corpus segments falling within the relevant publisher's catalogue.

Publisher TDM policies may exist in machine-readable form — for example through robots.txt directives, rights-retention metadata embedded in article-level records, or structured opt-out signals in metadata feeds — or in human-readable form as standalone policy documents published on the publisher's website or within licence agreements. Both formats carry legal and operational significance and must be actively monitored, as publishers may update their policies over time in response to developments in AI regulation or commercial negotiation. The dataset pipeline should include a mechanism for version-tracking policy documents and recording the date of each verification exercise. Critically, the absence of an explicit TDM policy on the part of a given publisher does not establish permission; in such cases, the applicable legal position must be assessed with reference to the relevant statutory framework — in particular CDSM Article 4 and its conditions — and any opt-out signals that may have been issued through other channels must be identified and respected.

11.2.3 Verification of Authorised-User Scope and Permitted TDM Activities

Institutional licence agreements customarily define the population of persons entitled to access the licensed content through the concept of an "authorised user," a category that typically encompasses human researchers, faculty members, and students affiliated with the subscribing institution. The question of whether automated, machine-driven access for the purposes of corpus extraction, tokenisation, or ingestion into an AI training pipeline falls within this designation is a matter that must be explicitly verified in each agreement and is not to be assumed. In the majority of legacy licence agreements — those concluded prior to the widespread deployment of AI training pipelines — such activities were neither contemplated nor addressed, and the concept of authorised use was formulated exclusively with human reading activity in mind. The extension of authorised-user rights to cover computational processes acting on behalf of an institution therefore requires either express contractual confirmation or a defensible legal basis under applicable statute.

Where reliance is placed on the non-commercial TDM exception provided under Article 4 of the CDSM Directive, the conditions attaching to that exception must be carefully mapped to the actual use case. The exception permits text and data mining of content to which the organisation has lawful access, provided that the use is non-commercial in nature, that the organisation falls within the category of permitted users, and that no opt-out has been exercised by the rightsholder in an appropriate and machine-readable manner. Where any of these conditions is not met — including where the project has a commercial dimension, where the user category is

uncertain, or where a publisher opt-out is in effect — the statutory exception cannot be relied upon, and additional authorisation is required.

11.2.4 Technical and Contractual Access Requirements

Independently of the question of substantive permission, many publishers impose specific technical and contractual prerequisites as a condition of bulk or programmatic content access for TDM purposes. These prerequisites commonly include the provision of a dedicated API key, the execution of a data-access agreement supplemental to the standard subscription terms, or both. The API key serves not only as a technical authentication mechanism but also as a contractual trigger, in the sense that its issuance is conditional upon acceptance of specific terms governing the permissible scope of data access, the format and volume of extraction, and the purposes for which extracted content may be used. Accessing publisher content through means other than the designated API channel — for instance, through automated scraping of the publisher's web interface — will in all likelihood constitute a breach of the subscription agreement, regardless of the substantive legitimacy of the intended use.

Where an institutional licence is silent on TDM or AI-training activities, this silence will typically necessitate the negotiation of an add-on or supplemental licence before corpus extraction is initiated. Proceeding with extraction in the absence of such an instrument, on the basis that the standard subscription terms do not expressly prohibit the activity, constitutes a licence breach and exposes the organisation to legal and financial risk. Contract review should therefore verify not only whether TDM activities are permitted but also whether separate written authorisation in the form of a data licence agreement or API Terms of Service is required as an additional condition of any such permission.

11.2.5 Heightened Risk Assessment for Commercial AI Development

The legal risk profile of LLM training on subscription-journal corpora is materially elevated where the model under development is intended for commercial deployment, whether as a standalone product, an integrated platform feature, or a general-purpose AI system. The research-oriented TDM exception established under Article 3 of the CDSM Directive — which applies to research organisations and cultural heritage institutions acting for the purposes of scientific research — does not extend to for-profit activities and is unavailable to commercial entities or to projects with commercial objectives. Reliance on this exception by entities or projects falling outside its subject-matter scope carries significant legal exposure.

The residual exception under Article 4, whilst broader in personal scope insofar as it is not confined to research organisations, is subject to the overriding condition that rightsholders have not exercised an opt-out in respect of the relevant content. In commercial contexts, where publishers have both the incentive and the mechanism to opt out of AI-training uses of their catalogues, the practical availability of Article 4 protection cannot be assumed and must be verified on a publisher-by-publisher and corpus-segment-by-corpus-segment basis. Commercial projects must therefore conduct a heightened risk assessment that maps each distinct component of the training corpus to its licensing status, identifies all applicable publisher opt-outs or contractual restrictions, quantifies the scale of any uncleared use, and documents the residual legal exposure in a form that informs management decision-making and, where

applicable, legal indemnity arrangements. This assessment should be completed and approved prior to the commencement of any training run on subscription-sourced content.

11.2.6 Requirement for an Explicit TDM or AI-Training Licence in Commercial Contexts

Where the AI development project is directed at a commercial product, a commercialised fine-tuned model, or a general-purpose AI system within the meaning of the EU AI Act, reliance on research-exception carve-outs is in principle insufficient as a basis for the use of subscription-journal content as training data. In such cases, the appropriate instrument is an explicit, purpose-specific TDM or AI-training licence negotiated directly with each relevant rightsholder or through a collective licensing mechanism where available. Such a licence must define with precision the scope of rights granted — covering reproduction for the purposes of tokenisation and model training, the creation of derivative model outputs, and any intended sublicensing of the trained model — as well as the territorial coverage of the grant, the duration of the licence, the categories of content covered, and any post-training obligations concerning retention or deletion of training corpora.

Licence negotiations should address, as a matter of priority, the question of downstream restrictions on model-sharing, API access to fine-tuned models, and dataset publication, given that publishers typically impose use-specific conditions that may constrain the commercialisation strategy if not identified and resolved at the negotiation stage. Where negotiations are protracted or where licences cannot be obtained for particular corpus segments, those segments should be excluded from the training corpus until appropriate authorisation is in place, and their exclusion should be documented in the project's compliance record.

11.2.7 Systematic Logging of Publisher Restrictions in the Dataset Register

Publisher TDM policies and licence agreements regularly impose use-specific restrictions that extend beyond the initial authorisation of corpus access and govern the downstream handling and deployment of content derived from licensed sources. Commonly encountered restrictions include prohibitions on the redistribution of full-text corpora, limitations on the sharing of derived or synthetic datasets generated from licensed content, and constraints on the publication or licensing of AI models trained on the relevant corpus. These restrictions are legally binding and operationally significant: their non-observance constitutes a breach of the applicable licence and may expose the organisation to enforcement action, injunctive relief, or financial liability.

To manage these obligations systematically, each licensed source incorporated into the training corpus should be assigned a corresponding entry in the organisation's dataset register. That entry must record, as a minimum: the relevant licence or policy reference number and version, the access method used for extraction, the nature and scope of all applicable downstream restrictions, and the date on which the licence terms were verified. This information must remain accessible and traceable throughout the AI development lifecycle — from corpus construction through model training, validation, and deployment — to ensure that licensing constraints are applied at every stage at which they are relevant and that the organisation is in a position to demonstrate compliance in the event of a regulatory inquiry or rights-holder challenge.

11.2.8 Operationalisation of No-Redistribution and Limited-Sharing Conditions

The contractual obligations arising from no-redistribution and limited-sharing clauses must be translated into concrete technical controls at the level of the data pipeline and the model deployment infrastructure, rather than managed solely as documentary compliance obligations. Full-text corpora, training dumps, and derivative datasets that are subject to redistribution restrictions must be stored in access-controlled environments, excluded from any external-facing data publication or sharing mechanism, and subject to data-governance policies that prevent inadvertent disclosure. The technical implementation of these controls must be verified before the corpus enters production use and must be maintained throughout the model training and deployment cycle.

Downstream product-design and commercialisation decisions — including decisions concerning model-sharing arrangements, the provision of API access to fine-tuned models trained on restricted content, and the publication of training dataset documentation — must be evaluated against the publisher-specific sharing restrictions recorded in the dataset register at the corpus-intake stage. Where a proposed deployment configuration would be inconsistent with an applicable restriction, deployment must be constrained, redesigned, or deferred until a compliant configuration has been established. Internal access controls, data-governance procedures, and contractual provisions governing third-party access must collectively ensure that licensed content does not propagate beyond the scope of the original authorisation at any point in the development and deployment lifecycle.

11.3 Abstracts, metadata, tables, and figures scenario

11.3.1 Copyright and Database Rights in Bibliographic Metadata

A common assumption in corpus construction is that restricting data ingestion to non-full-text content — such as titles, abstracts, and structured bibliographic metadata — eliminates or substantially reduces intellectual-property obligations. This assumption is legally unsound. Titles, abstracts, and bibliographic fields remain subject to copyright protection under general principles of EU copyright law insofar as they constitute the author's own intellectual creation, and are additionally protected, in the EU, by the sui generis database right established under Directive 96/9/EC of the European Parliament and of the Council. The sui generis right attaches to the maker of a database that demonstrates substantial investment in obtaining, verifying, or presenting its contents, irrespective of whether the individual records contained therein are themselves protected by copyright.

At the level of individual elements, certain metadata components — such as digital object identifiers (DOIs), publication dates, and author names — may not individually satisfy the originality threshold required for copyright protection, nor meet the qualitative or quantitative substantiality threshold of the sui generis right. However, this analysis cannot be applied globally to the corpus as a whole. The systematic compilation of such elements in a structured, machine-readable dataset is precisely the type of activity that database rights are designed to govern. The extraction or re-utilisation of a substantial part of a protected database — assessed cumulatively across repeated or systematic operations — constitutes an infringement of the sui generis right regardless of whether the individual records extracted are themselves unprotected. Organisations constructing training corpora from metadata aggregators, citation databases, or

bibliographic repositories must therefore conduct an independent rights analysis for each data layer, and must not rely on the reduced expressive length of metadata as a proxy for the absence of intellectual-property obligations.

11.3.2 Applicability of TDM Exceptions and Licensing Frameworks

The text and data mining (TDM) exceptions introduced by Articles 3 and 4 of Directive (EU) 2019/790 on Copyright in the Digital Single Market (the CDSM Directive) apply equally to corpora composed exclusively of metadata, abstracts, and structured supplementary data. The eligibility criteria governing each exception are not relaxed by reason of the reduced volume or granularity of the content in question. Under Article 3, which provides a mandatory exception for TDM carried out for the purposes of scientific research, the reproducing party must demonstrate lawful access to the source material and must act within a non-commercial research context; these conditions apply irrespective of whether the object of mining is a full-text article or an abstract dataset. Under Article 4, which provides a general exception subject to the rightsholder's opt-out right, compliance with any machine-readable reservation of rights expressed by the rights holder or by the database operator is obligatory, and the absence of opt-out indicators must be affirmatively verified prior to ingestion rather than presumed.

Publisher-specific TDM policies and the terms of use associated with bibliographic APIs and citation databases must be consulted and documented prior to bulk acquisition. Such policies frequently distinguish between modalities of machine-readable access — for instance, differentiating between access via a designated TDM API and programmatic access via a standard search or discovery interface — and the permissible uses under each modality may differ substantially. Where the chosen legal basis is a contractual licence rather than a statutory TDM exception, the standard licence-variant analysis applicable to open-access content must be applied without modification: Creative Commons licences permitting adaptation and commercial use (CC BY, CC BY-SA) must be distinguished from those imposing non-commercial or no-derivatives restrictions (CC BY-NC, CC BY-ND), and only licence variants that expressly or constructively permit AI training and computational reproduction should be accepted as a sufficient legal basis for inclusion in a training corpus.

11.3.3 Preferred Access Pathways and Open Bibliographic Sources

In order to minimise legal uncertainty in corpus construction, priority should be given to bibliographic data sources distributed under open licences with broad data-use permissions, in particular those released under the Creative Commons CC0 public domain dedication or under CC BY licences that do not restrict computational reuse. Several large-scale bibliographic infrastructure projects — including open metadata repositories and persistent identifier registries — provide bulk-download mechanisms or APIs under explicit data-use agreements that expressly permit AI training and computational analysis; these should be preferred over general-purpose data sources that lack equivalent clarity. Where such openly licensed alternatives are not available or are insufficiently comprehensive for the research purpose, official bulk-download mechanisms and publisher APIs should be used in preference to ad hoc web scraping. Access via designated machine-readable TDM pathways typically carries distinct and more permissive terms than general-purpose access, and the use of scraping tools on publisher websites may constitute a violation of terms of service that operates independently of copyright and database-right considerations.

11.3.4 Residual Copyright and Database-Right Risk in Metadata-Only Corpora

Reliance on metadata and abstracts in lieu of full-text content reduces, but does not eliminate, copyright and database-right risk. The reduced probability of verbatim regurgitation of extended protected passages during model inference is a relevant risk-mitigation factor, but it does not constitute a legal safe harbour, and it does not bear on the question of whether infringing copying occurred during the training phase itself. Training a model on reproduced metadata or abstract text involves the making of intermediate copies, which remains a restricted act under copyright law and which requires either a statutory exception or a licence as its legal basis, regardless of whether the outputs subsequently produced by the model reproduce any protected expression.

Furthermore, the systematic extraction and integration of bibliographic metadata at the scale required for LLM training corpus construction is, in itself, the type of activity most directly targeted by the sui generis database right. Even where individual records are not protected, and even where the reproducing party has lawful access to the source database, the cumulative extraction of substantial portions of the database's contents — whether in a single operation or through repeated smaller extractions — may constitute a prohibited re-utilisation under Article 7 of Directive 96/9/EC. This risk is heightened where the corpus is derived from commercially operated citation or abstract databases, which typically represent the beneficiaries of substantial investment in data collection and verification.

11.3.5 Privacy Obligations Arising from Bibliographic Metadata

Bibliographic metadata cannot be treated as inherently impersonal data. Metadata records in scientific literature databases frequently contain directly identifiable personal data — including the full names of authors, their institutional affiliations at the time of publication, acknowledgement sections identifying funders and collaborators, and supplementary contributor information — all of which constitute personal data within the meaning of Article 4(1) of Regulation (EU) 2016/679 (the General Data Protection Regulation, GDPR) where the data subjects are identifiable natural persons. The processing of such data in the course of training corpus construction — including its collection, storage, organisation, and use for model training — constitutes processing within the meaning of Article 4(2) GDPR, and requires a valid legal basis under Article 6 of that Regulation. Where the processing serves scientific research purposes, Article 89 GDPR and the relevant implementing legislation of EU Member States may provide derogations from certain data subject rights, but these derogations are conditional and must be assessed on a case-by-case basis.

A residual re-identification risk must be assessed even for metadata fields that appear to be anonymised or aggregated. The combination of bibliographic metadata with external datasets — such as institutional directories, author-identifier registries, or citation networks — may enable linkage attacks that restore the identifiability of data subjects in ostensibly non-identifying records. This risk is not merely theoretical; it is a recognised vulnerability in the scientific metadata ecosystem, given the public accessibility of multiple overlapping data sources. The privacy impact assessment required for high-risk processing operations under Article 35 GDPR should therefore extend to any metadata corpus of substantial scale, and should specifically address the risk of re-identification through dataset combination.

11.3.6 Provenance Documentation Requirements

The obligation to maintain comprehensive corpus-level provenance documentation applies to metadata and abstract datasets with the same rigour required for full-text corpora. For each data source or dataset segment included in the training corpus, provenance records must capture, at minimum: the source URL, API endpoint, or data repository from which the data was obtained; the date or date range of acquisition; the legal basis for inclusion, specifying whether the applicable basis is a statutory TDM exception (and if so, which provision of the CDSM Directive or equivalent national law), a contractual licence (with the specific licence variant identified), or a public domain or CC0 dedication; any opt-out indicators encountered and the disposition applied; and a description of any filtering, deduplication, or pre-processing operations applied to the data prior to ingestion.

This documentation obligation is reinforced by the transparency requirements imposed on providers of general-purpose AI (GPAI) models under Article 53(1)(d) of Regulation (EU) 2024/1689 (the EU AI Act). That provision requires GPAI providers to make available a sufficiently detailed summary of the content used for training, enabling copyright holders to exercise the rights conferred upon them by EU law. The EU AI Act does not limit this obligation to full-text training data; the summary must cover training content irrespective of the level of abstraction or granularity at which it was ingested. Provenance documentation must therefore be maintained in a form that is sufficiently detailed to support the production of such summaries, and must be preserved throughout the lifecycle of the model for which the training corpus was constructed.

11.3.7 Adversarial Testing and Leakage Mitigation

The restriction of training data to metadata and abstracts reduces, but does not preclude, the risk of model outputs reproducing protected expression or disclosing sensitive personal data. Red-teaming and adversarial testing exercises must therefore be conducted on models trained on such corpora in order to probe whether the model is capable of reproducing substantial portions of source works — including, for instance, extended abstract text that may itself constitute protected expression — or of disclosing personally identifiable information embedded in metadata fields. Copyright-focused adversarial testing should be designed to identify instances in which the model reproduces text strings of sufficient length and specificity to raise an inference of memorisation, while privacy-focused testing should include attempts to elicit disclosures of author names, institutional affiliations, or other personal data that may have been retained from the training corpus.

Such testing should be integrated into the pre-deployment validation pipeline as a mandatory step, and should be repeated at defined intervals following deployment, particularly following model updates or fine-tuning operations. Where leakage is detected, remediation measures must be applied proportionate to the nature and severity of the risk identified; such measures may include output-level filtering, the exclusion of implicated data records from future training iterations, the implementation of deployment restrictions, or, where the leakage is systemic, the retraining of the model on a revised corpus. The results of red-teaming exercises and any remediation actions taken should be recorded in the project's compliance documentation to support audit and regulatory review.

11.4 Consortium-shared material scenario

The integration of corpora contributed by multiple consortium partners into a shared AI training pipeline introduces a layered set of legal, contractual, and governance obligations that extend well beyond those applicable to datasets acquired by a single entity. Each corpus that enters the pipeline via a partner brings with it its own acquisition history, its own legal basis, and its own set of attached constraints — all of which must be verified, documented, and enforced by every receiving party. The following subsections set out the principal compliance requirements applicable to this scenario, addressing legal basis verification, licence analysis, data classification, inherited restrictions, filtering obligations, personal data exposure, provenance traceability, sharing constraints, and pre-redistribution governance.

11.4.1 Legal Basis and Verification of Contributed Corpora

Each dataset or corpus contributed by a consortium partner must be accompanied by documented evidence establishing the legal basis under which it was originally acquired and under which its downstream use for AI model training is authorised. Depending on the nature of the material and the acquiring entity, the applicable basis may be the text and data mining (TDM) exception for scientific research organisations under Article 3 of Directive (EU) 2019/790 on copyright in the Digital Single Market (CDSM Directive), the general TDM exception under Article 4 of the same Directive, a Creative Commons or bespoke contractual licence, or public-domain status established through expiry of copyright protection.

Partners acting as data contributors occupy the role of primary controllers or licensees with respect to their respective corpora. Notwithstanding this, the receiving entity must independently verify that the declared legal basis is valid and that it is applicable to the intended downstream use, namely the training of large language models (LLMs) on medical scientific literature. Reliance on the contributing partner's self-declaration is insufficient from a compliance standpoint: the receiving entity must examine the underlying documentation, assess the compatibility of the declared basis with the receiving organisation's own legal profile, and confirm that the intended use falls within the scope of the authorised purpose.

Of particular significance in consortium contexts is the scope of the Article 3 CDSM research-organisation exemption. That provision grants a mandatory exception to the reproduction rights of rightholders in favour of research organisations and cultural heritage institutions carrying out TDM for the purposes of scientific research. Where a contributing partner has relied on this exemption as the basis for acquiring and processing published scientific literature, the consortium must confirm that the exemption extends to — or is at minimum compatible with — the receiving partner's organisational classification and the commercial or non-commercial nature of the intended training activity. Where any doubt subsists as to this compatibility, the corpus must not be incorporated into the shared training pipeline until a legal assessment has resolved the question in writing.

11.4.2 Licence Analysis and Reuse Permissions

Prior to accepting shared material into the training pipeline, the receiving partner is required to obtain a precise mapping of all applicable licence variants governing each sub-corpus. This

mapping must distinguish, at minimum, between Creative Commons licences that are broadly permissive with respect to derivative works (CC BY and CC BY-SA), those that restrict use to non-commercial purposes (CC BY-NC and CC BY-NC-SA), and those that prohibit the creation of derivative works (CC BY-ND and CC BY-NC-ND). The latter category is in many jurisdictions incompatible with the use of the licensed material as an input to AI model training, given that such training constitutes a form of derivative processing that generates new outputs informed by the ingested content.

Any instance of ambiguous or undocumented licence information must be treated as a blocking condition. Material whose licence status cannot be confirmed with reasonable certainty must be excluded from the training corpus until the status is resolved and the resolution is recorded in the corpus documentation. This requirement is not susceptible to relaxation on grounds of practical convenience or timeline constraints, as integration of insufficiently licensed material into a training pipeline may expose all consortium partners to copyright infringement liability.

The consortium must additionally establish a shared governance agreement that specifies the conditions under which partner-contributed material may be used across the partnership. This agreement must address, inter alia, whether attribution requirements attached to source licences survive the contribution and are binding on all receiving parties, whether ShareAlike propagation obligations apply to outputs generated from the shared corpus, and whether any bespoke licence obligations negotiated by a contributing partner with specific rightsholders are enforceable downstream.

11.4.3 Classification of Material Categories

Published scientific literature — encompassing peer-reviewed journal articles, conference proceedings, and preprint deposits — and partner-derived corpora — such as internal institutional datasets, curated annotated collections, or proprietary experimental data — are governed by distinct legal regimes and must therefore be classified and maintained as separate categories within the data pipeline architecture.

Published literature primarily engages copyright law and the applicable TDM exceptions under the CDSM Directive, together with any publisher-imposed contractual conditions attaching to licences or API-based access. Partner-derived corpora, by contrast, may additionally engage contractual obligations arising from the terms under which the data were originally collected or commissioned, trade-secret protections applicable to proprietary methodologies or results, and — where the corpus contains individual-level data — the full data protection framework established by Regulation (EU) 2016/679 (GDPR) and, where applicable, sector-specific health-data frameworks such as those established under national transpositions of the Clinical Trials Regulation or hospital-data access schemes.

Technical and governance controls must reflect this classification distinction. Each category must be maintained in a separate provenance register, subject to distinct retention and access policies, and assessed through differentiated risk frameworks appropriate to the specific legal regime governing it. The conflation of these two categories — whether in storage architecture, governance documentation, or compliance assessments — is liable to produce systematic compliance gaps that may not be identified until audit or enforcement proceedings.

11.4.4 Inherited Contractual Restrictions and Source-Specific Constraints

Contractual restrictions attaching to the original source of contributed material travel with that material through the supply chain and are binding on all downstream consortium partners unless they have been explicitly released in writing by the relevant rightsholder or licensor. Such restrictions may originate from publisher-specific TDM policy conditions — including prohibitions on the redistribution of full-text corpora or on the sharing of trained model weights derived from licensed content — from institutional licence agreements governing access to bibliographic databases, from API terms of service imposing rate limits and redistribution bans, or from bespoke data-sharing agreements concluded at the point of original acquisition.

The contributing partner is required to provide the receiving partner with a complete catalogue of all source-specific constraints attaching to the contributed corpus. This catalogue must identify, with sufficient precision to enable downstream enforcement: any prohibitions on corpus redistribution within or outside the consortium; any field-of-use limitations restricting the purposes for which the data may be processed; any conditions governing the disclosure of trained models or derived datasets to third parties; and any time-limited or geographically scoped restrictions on use. The receiving partner must incorporate these constraints into its own product and deployment design and must ensure that they are reflected in any sub-agreements concluded within the consortium.

Where uncertainty arises as to whether a specific restriction applies to consortium redistribution or to a particular downstream deployment context, that uncertainty must be resolved through direct clarification with the relevant rightsholder or through the negotiation of an explicit TDM or AI-training add-on licence before the material is integrated into the shared pipeline. Proceeding in the face of unresolved ambiguity is not a permissible risk management strategy in this context.

11.4.5 Upstream Filtering, Opt-Out Compliance, and Exclusion Records

Under Article 4 of the CDSM Directive, rightholders retain the right to opt out of the general TDM exception by signalling a reservation of rights in an appropriate manner — for online content, through machine-readable mechanisms such as robots.txt directives or publisher-level opt-out metadata. The contributing partner bears primary responsibility for detecting and honouring these opt-out signals prior to contributing the corpus to the consortium. This obligation cannot be discharged by delegation alone: the receiving partner must also apply independent verification and must not assume that contributed corpora have been filtered without examining the associated documentation.

Documentation confirming that opted-out works, content licensed exclusively under ND or NC-ND variants incompatible with derivative training uses, and blacklisted sources have been excluded from the contributed corpus must accompany any data transfer. Where such filtering has been applied, a structured exclusion log must be produced and transmitted with the corpus. This log must record, for each excluded item or source: the nature of the opt-out or exclusion signal that triggered the removal; the identity of the excluded works or source collections; and the date on which the exclusion was applied. The log must be incorporated into the corpus provenance record and must be preserved in a format suitable for audit, including potential regulatory audit in the context of EU AI Act compliance or copyright enforcement proceedings.

11.4.6 Personal and Sensitive Data in Shared Corpora

Prior to transfer, the contributing partner is required to perform a systematic scan of the corpus to identify any personal data within the meaning of Article 4(1) GDPR, including direct identifiers such as author names, patient identifiers appearing in case reports or clinical supplements, and indirect identifiers such as acknowledgement sections, qualitative interview transcripts, or individual-level supplementary datasets. The identification of personal data in a corpus intended for LLM training triggers the full GDPR compliance framework, including the requirement to establish a lawful basis for processing, to assess the compatibility of the training use with the original purpose for which the data were collected, and to implement the safeguards required under Article 89(1) for research purposes.

Where special-category personal data within the meaning of Article 9 GDPR are detected or reasonably suspected to be present — including health data, genetic data, or biometric data in the context of medical scientific literature — the contributed corpus must be subject to a heightened data protection review before transfer. This review must confirm the existence of an explicit lawful basis under Article 9(2), verify that the applicable Article 89 safeguards are in place, and — where required by national law or by the terms of the ethics-committee approval under which the data were originally collected — obtain the authorisation of the relevant ethics committee or competent health-data authority for the AI training use.

The receiving partner must not assume that de-identification or anonymisation applied by the contributing partner is sufficient to remove personal data from the scope of GDPR. The risk of re-identification — particularly when combined with other corpora held by consortium partners, or when the trained model retains memorised fragments of training data — must be independently assessed by the receiving partner before integration into the training pipeline. This assessment should be documented as part of the Data Protection Impact Assessment (DPIA) required for high-risk processing operations under Article 35 GDPR.

11.4.7 Provenance Traceability and Transformation Logging

Each contributed corpus must be accompanied by a machine-readable provenance record capturing, at minimum: the original source URLs, database identifiers, or API endpoints from which the content was acquired; the method and date of acquisition; the copyright licence or TDM exception relied upon at the time of acquisition; any applicable opt-out indicators or exclusion flags; and the identity of the consortium partner responsible for the acquisition. This record must be structured so as to enable automated integrity checking against known catalogues of published scientific literature, supporting copyright auditing and the identification of inadvertently included prohibited material.

All transformation steps applied to the corpus prior to its contribution to the consortium must be logged with sufficient granularity to permit reproduction of the final corpus state. Transformation steps subject to this logging requirement include deduplication, anonymisation and pseudonymisation, format conversion, textual filtering, and statistical sampling. Each log entry must record the transformation type, the parameters applied, the tool or script used, and the date on which the transformation was executed. The transformation log forms part of the corpus provenance record and must be maintained with the same retention and access standards.

Governance and approval history must be recorded in a living document linked to each corpus entry in the data register. This includes documentation of internal governance body sign-off on inclusion of the corpus in the training pipeline, the outcomes of legal and intellectual property review, and confirmation of DPIA completion where required. The living document must be updated whenever the corpus itself is modified, or whenever its intended use changes — for example, if a corpus approved for use in a non-commercial internal research tool is subsequently considered for deployment in a commercially licensed clinical decision-support application.

11.4.8 Sharing Constraints, Permitted Uses, and Deployment Scope

The consortium must maintain a centralised dataset register documenting, for each shared corpus, the full scope of permitted uses, approved deployment contexts, and applicable constraints on redistribution. The register must specify, in terms sufficiently precise to guide product and system design: whether the corpus may be used exclusively for non-commercial research training or also for commercially deployed models; whether the trained model may be deployed in internal research environments only or may also be embedded in externally accessible clinical tools; and whether redistribution of trained model weights or derived datasets to parties outside the consortium is prohibited under applicable source licences or contractual terms.

Constraints inherited from source licences, publisher terms, and GDPR-based approvals must be translated into enforceable downstream obligations through contractual instruments between consortium partners. Where any consortium partner engages a third-party processor to carry out training operations on shared corpora, a Data Processing Agreement (DPA) compliant with Article 28 GDPR must be concluded prior to transfer of personal-data-containing corpora. The DPA must give effect to all applicable constraints derived from the corpus provenance record.

The dataset register and associated sharing-constraint documentation must be subject to version control and must be updated promptly to reflect changes in applicable law, regulatory guidance issued by competent supervisory authorities, or modifications to rightsholder policies. Of particular relevance in the current regulatory environment are the evolving implementation guidance issued under the EU AI Act and new enforcement positions adopted by data protection authorities with respect to the use of personal data in AI training — both of which may alter the permissibility of uses previously considered compliant.

11.4.9 Governance Validation and Pre-Redistribution Authorisation

Before any shared corpus is redistributed within the consortium for training or fine-tuning purposes, a dedicated governance checkpoint must be completed. This checkpoint must encompass a copyright and licence audit conducted and signed off by the consortium's Legal and Intellectual Property function; a GDPR compliance review completed by the Data Protection function, including confirmation of DPIA adequacy where applicable; and a risk classification assessment conducted by the AI Governance function in accordance with the risk categories established under the EU AI Act and the consortium's internal AI governance framework.

The governance body or consortium steering committee must issue a documented go/no-go decision before training commences on any shared corpus. This decision must be recorded in writing and must identify any conditions, constraints, or ongoing monitoring requirements

attached to the approval. Where approval is conditional, the conditions must be enforceable and their satisfaction must be verified prior to commencement of training operations. Undocumented oral approvals or implicit authorisations by inaction are not permissible in this governance context.

Ongoing compliance of shared training datasets with copyright, licensing, and data-protection obligations must be assessed through regular internal or third-party audits at intervals appropriate to the sensitivity of the material and the pace of regulatory development. Post-training compliance activities must include red-teaming exercises focused on the identification of memorisation artefacts and sensitive-data leakage in model outputs. The results of such exercises must be fed back into the governance process and must, where necessary, trigger remediation measures including targeted unlearning procedures, data exclusion from subsequent training runs, or revision of the corpus provenance record.

12 Conclusions

The analysis undertaken in this deliverable reveals that the use of medical scientific literature as training data for large language models operates within a legal environment of considerable structural complexity, in which no single regulatory instrument provides a self-sufficient basis for compliance. The applicable framework is constituted by the concurrent and often tensional interaction of copyright law, the sui generis database right, the General Data Protection Regulation, the EU AI Act, and the contractual regimes imposed by publishers, licensing bodies, and data infrastructure providers. The central finding of this guide is that lawful and responsible use of scientific literature for AI training cannot be reduced to a binary assessment of copyright permissibility; it demands a layered, source-level analysis that simultaneously addresses the legality of access, the scope of applicable exceptions, the enforceability of contractual restrictions, the presence and treatment of personal data, and the propagation of all such constraints through the full data lifecycle and into downstream model deployment.

The text and data mining exceptions established under Articles 3 and 4 of the CDSM Directive, while constituting the principal statutory basis for lawful corpus construction in the European Union, are subject to limitations that significantly constrain their practical utility for large-scale AI training. Article 3 offers the most robust legal foundation, as its mandatory and contractually non-overridable character shields qualifying research organisations from publisher restrictions and opt-out mechanisms. However, this exception is strictly delimited by a double condition: only entities meeting the Directive's definition of research organisations or cultural heritage institutions may invoke it, and only for activities whose purpose is genuinely scientific and non-commercial. Any commercial dimension in the training activity, any involvement of a consortium partner that does not satisfy the beneficiary condition, or any downstream deployment of the trained model for revenue-generating purposes risks displacing the entire legal basis from Article 3 to Article 4, where the rightsholder opt-out mechanism, contractual restrictions, and publisher licensing terms become fully operative. This transition from a mandatory to a qualified exception fundamentally alters the compliance posture of the project and creates a latent legal fragility that must be managed through continuous purpose-classification and institutional-status verification.

The practical effect of Article 4's opt-out regime, reinforced by the transparency and copyright-compliance obligations imposed on providers of general-purpose AI models under Article 53 of the EU AI Act, is that the overwhelming majority of subscription-based scientific literature is now effectively foreclosed for commercial AI training absent specific contractual authorisation. Major academic publishers have systematically deployed machine-readable opt-out signals, incorporated AI-training prohibitions into their terms of service, and developed parallel commercial licensing programmes that monetise their content holdings as training data. The emergence of this licensing market, already valued in the hundreds of millions of dollars, establishes a commercial expectation that courts and regulators are likely to consider when assessing whether unauthorised training constitutes market harm. For the consortium, this means that reliance on statutory exceptions alone is an increasingly precarious strategy for any activity with a commercial horizon; licensing negotiations with publishers must be treated as a structural component of the project's data sourcing architecture rather than as a contingency measure.

The interaction between copyright, database rights, and contractual restrictions produces a compliance environment in which rights are stacked rather than alternative. A single act of large-scale extraction from a publisher platform may simultaneously engage article-level copyright, platform-level sui generis database rights, subscription-agreement restrictions, API terms of service, and machine-readable TDM opt-outs. The legal analysis required to clear such an act cannot be conducted at the corpus level; it must proceed at source granularity, and in many cases at article or licence-variant level, given that open-access content published by the same publisher may carry different Creative Commons variants with materially different implications for commercial use. The operational consequence is that compliant corpus construction demands licence-aware ingestion pipelines capable of enforcing filtering, segmentation, and exclusion decisions at the point of data retrieval, not as a post-collection audit. Any architecture that defers rights verification to a later pipeline stage creates an interval of unquantified legal exposure during which infringing reproduction may already have occurred.

A further dimension of strategic significance is the analytical independence between the legality of training-stage activities and the legality of model outputs. The TDM exceptions authorise specific acts of reproduction and extraction for the purpose of analysis; they do not extend to the further distribution of mined works or to the generation of outputs that reproduce substantial portions of protected expression. This distinction means that a model trained on a lawfully constructed corpus may nonetheless generate infringing outputs if it memorises and reproduces recognisable passages from copyrighted works. The risk is particularly acute in the medical domain, where the training corpus may include paywalled or subscription-access articles and where prompted generation of verbatim content could constitute an unlicensed making available. The Hamburg Higher Regional Court's 2026 framework, which distinguishes dataset creation, model training, and output generation as three analytically independent phases each requiring its own legal basis, confirms that compliance must be addressed at every stage of the AI lifecycle and that lawfulness at one stage does not immunise subsequent stages from scrutiny.

The data protection dimension introduces constraints that operate independently of, and in parallel with, the intellectual property framework. The EDPB's Opinion 28/2024 has established that AI models trained on personal data cannot be presumed anonymous, that the legitimate interest basis for training is subject to a rigorous three-part cumulative test, and that development and deployment constitute distinct processing operations each requiring an

independent legal basis. For a project operating in the biomedical domain, where training corpora are likely to contain health data, genetic identifiers, clinical case details, and other special categories of personal data, the processing triggers the heightened protections of Article 9 GDPR in addition to the general data protection obligations. The scientific research exemption under Article 89 provides a pathway but imposes substantive conditions, including pseudonymisation, data minimisation, and documented safeguards, whose implementation must be demonstrated rather than merely asserted. The concurrent obligation under Article 10 of the EU AI Act to ensure that training datasets are sufficiently representative and complete creates a structural tension with the GDPR's minimisation principle: the consortium must navigate the narrow space in which both requirements are simultaneously satisfied, a task that demands documented justifications calibrated to the specific clinical populations and use cases for which the system is intended.

The regulatory landscape in which this project operates is not static. The European Parliament's March 2026 Resolution on Copyright and Generative AI signals a potential legislative shift towards mandatory licensing fees, expanded transparency requirements, and the extension of EU copyright law's territorial reach to cover training conducted outside the Union. The pending CJEU preliminary reference in Like Company v Google Ireland (Case C-250/25) will, for the first time, directly address whether LLM training on in-copyright content constitutes reproduction within the meaning of the InfoSoc Directive and whether the Article 4 TDM exception applies. The European Commission's proposed GDPR reform under the Digital Omnibus package, if adopted, would codify legitimate interest as a harmonised legal basis for AI model training while leaving the Article 9 safeguards for special categories of data intact. Each of these developments has the potential to materially alter the permissibility of practices currently considered compliant or to create new compliance obligations that must be addressed retrospectively. The consortium's governance framework must therefore be designed for adaptability, incorporating horizon-scanning mechanisms, periodic re-assessment triggers, and the capacity to modify or reconstruct training corpora in response to changes in the legal environment.

From a governance and operational standpoint, the analysis demonstrates that documentation and traceability are not ancillary administrative tasks but constitute the evidential foundation upon which the entire compliance architecture depends. The EU AI Act's transparency requirements for GPAI model providers, the CDSM Directive's reliance on demonstrated lawful access and opt-out compliance, and the GDPR's accountability principle all converge on the same operational imperative: the consortium must be capable of reconstructing, for any given model version, precisely which corpus content was used, under which legal basis, through which access pathway, subject to which licence conditions, and with which safeguards applied. A failure at any point in this documentation chain propagates upward through the compliance structure, rendering the entire legal position indefensible regardless of whether the underlying substantive compliance was in fact adequate at the time of processing. The governance architecture must therefore operationalise documentation as a prospective control embedded in pipeline design, not as a retrospective reporting obligation, and must assign clear ownership, escalation pathways, and periodic review cycles to ensure that governance decisions remain current and auditable.

The consortium-specific dimension of the project introduces additional layers of complexity that cannot be addressed through unilateral compliance strategies. Each partner brings its own institutional status, its own access rights, its own contractual relationships with publishers, and

its own national transposition of the applicable EU directives. The legal basis upon which one partner lawfully acquires and processes a corpus does not automatically extend to other partners, and the research-organisation status that entitles one entity to invoke the Article 3 TDM exception may not be shared by all members of the consortium. Contributed corpora carry inherited restrictions that are binding on all receiving parties, and the receiving entity bears an independent obligation to verify the declared legal basis, assess its compatibility with the intended downstream use, and confirm that no opt-out or contractual constraint has been overlooked. The shared governance agreement must establish clear rules for licence propagation, permissible use boundaries, data classification, and pre-redistribution authorisation, ensuring that compliance obligations are enforceable across institutional boundaries and that no partner's activities expose the consortium as a whole to unmitigated legal risk.

In light of the foregoing, the strategic orientation of the project should prioritise the construction of a legally resilient corpus architecture founded on three principles. First, maximise the use of openly licensed content released under CC0 or CC BY terms, which provides the highest degree of legal certainty and eliminates the need to invoke statutory exceptions or navigate publisher opt-out regimes. Second, maintain strict segregation between corpus partitions governed by different licensing regimes, enforcing use-case boundaries through technical controls that prevent cross-contamination between commercial and non-commercial pipelines. Third, embed compliance verification, rights resolution, and documentation requirements as structural components of the ingestion pipeline, ensuring that no content enters the training corpus without a positively confirmed and recorded legal basis. Where the corpus includes content for which the legal position is uncertain or evolving, the precautionary default must be exclusion pending resolution. This approach does not eliminate all legal risk, but it establishes a defensible compliance posture that is proportionate to the complexity of the regulatory environment and responsive to the legitimate expectations of rightsholders, data subjects, and regulatory authorities.

13 Bibliography

1. **Directive 2001/29/EC of the European Parliament and of the ... - WIPO**. Disponible en: https://www.wipo.int/wipolex/en/legislation/details/1453
2. **Directive No. 96/9/EC of the European Parliament and ...**. Disponible en: https://www.wipo.int/wipolex/en/legislation/details/1409
3. **Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (CDSM Directive)**. Disponible en: https://eur-lex.europa.eu/eli/dir/2019/790/oj
4. **EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models**. Disponible en: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf
5. **EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR**. Disponible en: https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf
6. **The General-Purpose AI Code of Practice**. Disponible en: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai
7. **MDCG 2025-6: Guidance on the interplay between the AI Act and the Medical Device Regulation / IVDR**. Disponible en: https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en
8. **European Parliament, AI and Copyright: Training of General-Purpose AI (policy brief)**. Disponible en: https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/769585/EPRS_ATA(2025)769585_EN.pdf
9. **UNESCO Recommendation on the Ethics of Artificial Intelligence**. Disponible en: https://www.ohchr.org/sites/default/files/2022-03/UNESCO.pdf
10. **OECD AI Principles**. Disponible en: https://www.oecd.org/en/topics/sub-issues/ai-principles.html
11. **Margoni, T. & Kretschmer, M., A Deeper Look into the EU Text and Data Mining Exceptions**. Disponible en: https://academic.oup.com/grurint/article/71/8/685/6650009
12. **The Text and Data Mining Opt-out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Suffocating Blanket for European AI Innovations?**. Disponible en: https://academic.oup.com/jiplp/article/19/5/453/7614898
13. **A literature review of "lawful" text and data mining**. Disponible en: https://open-research-europe.ec.europa.eu/articles/4-153
14. **Legal reform to enhance global text and data mining research**. Disponible en: https://www.science.org/doi/10.1126/science.add6124
15. **Text and Data Mining of In-Copyright Works: Is It Legal?**. Disponible en: https://cacm.acm.org/opinion/text-and-data-mining-of-in-copyright-works
16. **Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU**. Disponible en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3470653

17. **Copyright Exceptions and Fair Use Defences for AI Training Done for "Research" and "Learning," or the Inescapable Licensing Horizon**. Disponible en: https://www.cambridge.org/core/product/identifier/S1867299X25100354/type/journal_article

18. **Lucchi, N., ChatGPT: A Case Study on Copyright Challenges for Generative AI Systems**. Disponible en: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/CEDCE34DED599CC4EB201289BB161965/S1867299X23000594a.pdf

19. **Riccio, G.M., AI, Data Mining and Copyright Law: Remarks about Lawfulness and Efficient Choices**. Disponible en: https://ieeexplore.ieee.org/document/10569189

20. **Dornis, T.W. & Stober, S., Generative AI Training and Copyright Law**. Disponible en: https://arxiv.org/abs/2502.15858

21. **Creative Commons, Using CC-Licensed Works for AI Training**. Disponible en: https://creativecommons.org/using-cc-licensed-works-for-ai-training-2

22. **PMC Article Datasets**. Disponible en: https://pmc.ncbi.nlm.nih.gov/tools/textmining

23. **PMC Open Access Subset**. Disponible en: https://pmc.ncbi.nlm.nih.gov/tools/openftlist

24. **Elsevier Text and Data Mining (TDM) License**. Disponible en: https://www.elsevier.com/about/policies-and-standards/text-and-data-mining/license

25. **Text and data mining (TDM) policy and licence**. Disponible en: https://bmjgroup.com/text-and-data-mining-tdm-policy

26. **Springer Nature TDM Reservation Policy**. Disponible en: https://datasolutions.springernature.com/tdm-reservation-policy

27. **arXiv License Information**. Disponible en: https://info.arxiv.org/help/license/index.html

28. **Cranfield University Library, Publisher policies for TDM**. Disponible en: https://library.cranfield.ac.uk/text-and-data-mining/publisher-policies

29. **GA4GH DURI (Data Use and Researcher Identity) product line**. Disponible en: https://ga4gh-duri.github.io/categories/welcome.html

30. **Article 10: Data and Data Governance | EU Artificial Intelligence Act**. Disponible en: https://artificialintelligenceact.eu/article/10

31. **Baack, S. et al., Towards Best Practices for Open Datasets for LLM Training**. Disponible en: https://arxiv.org/abs/2501.08365

32. **Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI**. Disponible en: https://arxiv.org/abs/2204.01075

33. **Datasheets for Datasets**. Disponible en: https://arxiv.org/abs/1803.09010

34. **The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards**. Disponible en: https://arxiv.org/abs/1805.03677

35. **The Landscape of ML Documentation Tools**. Disponible en: https://huggingface.co/docs/hub/model-card-landscape-analysis

36. **Data Portraits: Recording Foundation Model Training Data**. Disponible en: https://arxiv.org/pdf/2303.03919.pdf

37. **Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments**. Disponible en: https://arxiv.org/pdf/2312.06153.pdf

38. **AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act**. Disponible en: https://arxiv.org/pdf/2406.18211.pdf

39. **SoK: Dataset Copyright Auditing in Machine Learning Systems**. Disponible en: https://arxiv.org/pdf/2410.16618.pdf

40. **Towards Data Governance of Frontier AI Models**. Disponible en: https://arxiv.org/pdf/2412.03824.pdf

41. **Improving governance outcomes through AI documentation: Bridging theory and practice**. Disponible en: https://arxiv.org/pdf/2409.08960.pdf

42. **Meszaros, J. & Ho, C.H., The Future Regulation of AI Systems in Healthcare**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC9576843

43. **Gallois, H. et al., Integrating AI into Health Care through Data Access**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC6813940

44. **Thorogood, A. et al., Purpose definition as a crucial step in the secondary use of personal data in scientific research under GDPR**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC10834358

45. **Raposo, V.L., Can personal data be recycled? The European Health Data Space and the GDPR**. Disponible en: https://academic.oup.com/ijlit/article/doi/10.1093/ijlit/eaae016/8157201

46. **Siebelmann, S. et al., Big Data-based Studies in Ophthalmology within the GDPR**. Disponible en: https://www.thieme-connect.de/DOI/DOI?10.1055/a-2165-9815

47. **Brauneck, A. et al., Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC10131784

48. **Addressing contemporary threats in anonymised healthcare data using privacy engineering**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC11885643

49. **The new EU–US data protection framework's implications for healthcare**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC11427690

50. **Publishers' Instructions on Use of Generative AI in Journal Submissions**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC10828852

51. **GA4GH standards enable the responsible sharing of genomic and health-related data**. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC9903747

52. **Using sensitive data to debias AI systems: Article 10(5) of the EU AI Act**. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/S026736492500010X