BOSONIT S. L.

Document detailing the assessment process, findings, and categorization strategy for the collected literature

DELIVERABLE 7.2

Table of contents

1 Introduction

The present document constitutes Deliverable 7.2 (D7.2) of the AIDESL project (Artificial Intelligence Data Extraction of Scientific Literature), an international research and innovation initiative co-funded under the ITEA Call 2023 programme. AIDESL brings together industry partners, research organisations, and academic institutions from Canada, Germany, Iceland, the Netherlands, Spain, and the United Kingdom, with the shared objective of applying large language models (LLMs) and other artificial intelligence techniques to fully automate the extraction of structured data from published medical and scientific literature, thereby reducing the time, cost, and error rates associated with the conduct of Systematic Literature Reviews (SLRs).

Within the project's architecture of deliverables, D7.2 corresponds to Task 7.3 (Data Assessment and Metadata Modelling) of Work Package 7 (Health Data Space Enrichment), led by Bosonit S.L. Work Package 7 addresses the foundational data management layer of the AIDESL platform, encompassing the legal and governance dimensions of data collection and processing, the assessment and categorisation of the collected literature, the design of data curation and storage infrastructure, the application of semantic enrichment techniques, and the integration of the resulting datasets into the project's federated health data space. D7.2 occupies a central position in this sequence: it builds upon the legal and compliance framework established in D7.1 and provides the methodological and analytical foundation upon which the curation, storage, and preprocessing activities described in D7.3 and D7.4 will be conducted.

This introduction is organised as follows. Section 1.1 describes the context and specific purpose of this deliverable within the broader AIDESL project. Section 1.2 maps its boundaries against the other deliverables produced under Work Package 7, with explicit reference to the content reserved for each. Section 1.3 provides a structural overview of the document.

1.1 Context and pupose of this deliverable

The ability to train large language models on high-quality, domain-specific scientific literature is a prerequisite for achieving the performance targets set out in the AIDESL Full Project Proposal. The quality of the resulting models is directly conditioned by the quality of the corpus on which they are trained: the relevance of the selected publications to the project's target use cases, the scientific rigour and methodological consistency of the included studies, the structural characteristics of the literature that determine its suitability for automated text processing, and the semantic coherence of the categorisation scheme applied to organise the corpus. These dimensions cannot be addressed through legal compliance criteria alone, nor through technical pipeline design in isolation. They require a dedicated analytical effort an informed, systematic, and documented evaluation of the collected literature that is the specific mandate of this deliverable.

D7.2 documents the process, findings, and outcomes of the data assessment and categorisation activities conducted by Bosonit S.L. under Task 7.3. Its purpose is threefold. First, it provides a transparent and auditable account of how the literature corpus assembled for AIDESL was evaluated, what criteria were applied, and what conclusions were drawn from that evaluation. Second, it defines the categorisation strategy adopted to organise the corpus in a manner that supports both LLM training and the semantic interoperability requirements of the project's federated health data space. Third, it describes the governance mechanisms applied during the assessment process to ensure that no article containing potentially sensitive information was admitted into the corpus without prior human review, in accordance with the data protection obligations established in D7.1.

The deliverable is the product of work conducted during the period from the project's inception in October 2024 through to the deliverable submission date of June 2026. During this period, the data assessment and categorisation activities evolved in parallel with the technical development of the data ingestion and processing pipeline, and the findings reported here reflect both the results of systematic evaluation and the methodological refinements introduced as the pipeline matured. Where applicable, the document distinguishes between assessment activities conducted on the initial corpus and those applied on an ongoing basis as new literature was ingested into the system.

The AIDESL project targets three primary application domains medical device development, pharmaceutical evaluation, and clinical research each of which places distinct demands on the corpus in terms of thematic coverage, evidence level, and document structure. The assessment and categorisation activities documented here were designed to address these domain-specific requirements explicitly, ensuring that the corpus supports not only general-purpose LLM training but also the specialised extraction tasks that define each of the project's demonstration scenarios. The findings reported in this deliverable therefore have direct operational consequences for the design of the preprocessing pipeline (D7.4), the structure of the data storage infrastructure (D7.3), and the selection and fine-tuning of AI and LLM techniques undertaken in Work Package 2.

The data assessment and categorisation methodology described in this document was developed and progressively refined through an iterative process informed by the empirical results obtained at each stage of the pipeline's development. The assessment framework was not defined a priori as a static specification but evolved in response to the characteristics of the literature encountered during ingestion, the performance feedback obtained from early LLM training experiments conducted within Work Package 2, and the evolving interoperability requirements of the federated health data space addressed in Work Package 7. This iterative approach reflects a deliberate methodological choice: given the heterogeneity of biomedical literature across the project's three target domains, a fixed assessment schema defined before any data had been processed would have been insufficiently sensitive to the structural and semantic variability of the corpus.

Finally, this deliverable contributes to the transparency and reproducibility obligations that govern AI development under the EU AI Act (Regulation (EU) 2024/1689). By documenting the composition, quality characteristics, and categorisation of the training corpus in a structured and publicly accessible format, D7.2 enables consortium partners and external stakeholders to understand the data foundations of the AIDESL platform, assess the representativeness and potential limitations of the corpus, and verify that the data governance principles established in D7.1 have been consistently applied throughout the assessment process.

1.2 Relationship to other WP7 deliverables

Work Package 7 produces six sequential deliverables, each addressing a distinct layer of the health data management and enrichment strategy pursued by Bosonit S.L. within AIDESL. The boundaries between these deliverables are defined with precision in order to ensure that each document addresses a well-delineated set of questions, that downstream outputs can build on the foundations established by their predecessors without duplication, and that the overall sequence of deliverables constitutes a coherent and internally consistent account of the work conducted under WP7. This section maps the specific scope of D7.2 against each of the remaining deliverables in the sequence, making explicit what this document covers, what it deliberately excludes, and where the reader should look for complementary content.

D7.1 Legal and IP Framework (delivered September 2025)

D7.1 established the comprehensive legal and intellectual property framework governing all data collection and processing activities conducted under WP7. It addressed, in full, the applicable regulatory instruments including the General Data Protection Regulation, the Directive on Copyright in the Digital Single Market, the EU AI Act, and the European Health Data Space Regulation as well as the source eligibility criteria, licensing and rights matrices, data lifecycle compliance requirements, consortium governance structure, and practical decision rules for recurrent data-use scenarios.

D7.2 builds directly upon the foundations laid in D7.1 but does not revisit them. The legal eligibility criteria defined in D7.1 are treated here as operative constraints that bound the scope of the assessment: only literature that satisfies the compliance requirements established in D7.1 is subject to the evaluation and categorisation activities documented in the present deliverable. Where this document makes reference to legal or governance considerations for instance, in the context of sensitive data detection during the assessment process it does so at the level of operational consequence rather than legal analysis, and the reader is directed to D7.1 for the underlying normative framework.

D7.3 Data Curation and Storage Infrastructure (due December 2026)

D7.3 will document the design and implementation of the data curation and storage infrastructure underpinning WP7, including the architecture of the storage system, the dataset documentation standards, the copyright clearance workflows at the operational level, and the catalogue of standardised datasets assembled for testing and validation purposes. The present document does not address storage architecture, infrastructure design, or operational curation workflows. Where the assessment and categorisation activities described in D7.2 generate outputs such as structured metadata records or categorised corpus partitions that are subsequently ingested by the storage infrastructure, this document describes those outputs at the level of their content and format, leaving the specification of the receiving infrastructure entirely to D7.3.

D7.4 Data Preprocessing for LLM Training (due March 2027)

D7.4 will provide a comprehensive guide to the data preprocessing pipeline applied to the AIDESL corpus in preparation for LLM training, covering text extraction techniques, data cleaning procedures, NLP-specific transformations, chunking strategies, parameterisation approaches, and the toolkits developed to operationalise these processes. The present document does not address preprocessing methodology or implementation. The categorisation and metadata outputs produced through the assessment activities described here serve as inputs to the preprocessing pipeline, and D7.2 documents those outputs in sufficient detail to enable D7.4 to build upon them; however, the technical specification of the preprocessing operations themselves is reserved entirely for D7.4.

D7.5 Health Data Space Architecture and Implementation Roadmap (due March 2027)

D7.5 will document the architecture of the federated health data space developed under WP7, including the interoperability standards adopted, the integration mechanisms with external data space connectors, and the implementation roadmap for scaling the solution beyond the project's current

scope. The present document does not address data space architecture or integration design. Where categorisation and semantic labelling decisions described in D7.2 have been informed by interoperability requirements for instance, the selection of MeSH as the primary controlled vocabulary in preference to SNOMED CT or HL7 FHIR the rationale is presented here in terms of its consequences for corpus organisation, with the broader architectural implications reserved for D7.5.

D7.6 Validation and Sustainability Report (due March 2027)

D7.6 will report on the validation of the outputs produced across WP7, including quality metrics for the assembled corpus, performance evaluations of the trained models, and an assessment of the long-term sustainability of the data management approach adopted by the project. The present document does not constitute a validation report. The assessment findings documented in D7.2 describe the characteristics of the corpus as observed through the evaluation process, but formal quality metrics, inter-annotator agreement scores, and model-performance-based corpus validation are reserved for D7.6.

The boundaries described above are summarised in the following table.

Deliverable	Primary focus	Content reserved for that deliverable	Content addressed in D7.2
D7.1	Legal and IP framework	Full legal analysis, compliance checklists, governance structure	Referenced as operative constraint; not reproduced
D7.3	Curation and storage infrastructure	Storage architecture, dataset catalogue, copyright clearance workflows	Assessment outputs described as inputs to D7.3
D7.4	Preprocessing for LLM training	Text extraction, NLP transformations, chunking, preprocessing toolkits	Categorisation outputs described as inputs to D7.4
D7.5	Health data space architecture	Interoperability design, data space integration, implementation roadmap	Vocabulary selection rationale at corpus-organisation level only
D7.6	Validation and sustainability	Formal quality metrics, model performance validation, sustainability assessment	Assessment findings as empirical observations; not formal validation

Table 1. Scope boundaries between D7.2 and other WP7 deliverables

1.3 Document structure

The present document is organised into seven sections, each addressing a distinct component of the data assessment and categorisation work conducted under Task 7.3 of Work Package 7. The following paragraphs describe the content and purpose of each section, enabling the reader to navigate the document efficiently and to locate specific topics without having to read the document in its entirety.

Section 2 provides an overview of the data collection process that preceded and informed the assessment activities documented in subsequent sections. It describes the repositories from which literature was sourced, the collection methodology applied, and the general scope and volume of the corpus assembled. This section does not revisit the legal eligibility criteria governing source selection which are established in full in D7.1 but situates the collection process within the operational context of WP7 and provides the factual basis upon which the assessment findings reported in Section 4 are grounded.

Section 3 sets out the assessment methodology developed and applied by Bosonit S.L. to evaluate the collected literature. It defines the assessment framework in terms of its three principal analytical dimensions relevance to the project's target use cases, scientific quality, and suitability for LLM training and describes the scoring approach, the decision rules applied at each evaluation stage, and the iterative refinements introduced to the methodology as the assessment process progressed. This section provides the methodological foundation necessary to interpret the findings reported in Section 4.

Section 4 presents the findings of the assessment process. It reports the overall characterisation of the corpus, disaggregated by repository of origin, by target application domain, and by document type, and identifies the principal gaps and limitations observed in the assembled literature. The findings reported in this section constitute the empirical basis for the categorisation strategy defined in Section 5 and provide the evidence base for the recommendations set out in Section 7.

Section 5 defines the categorisation strategy adopted to organise the assessed corpus. It presents the taxonomy of categories developed for AIDESL, describes the semantic labelling approach based on Medical Subject Headings (MeSH) controlled vocabulary, explains the methodological rationale for the selection of MeSH in preference to alternative biomedical terminology standards, and documents the categorisation process and the tools developed to implement it at scale. The section concludes with a quantitative description of the corpus distribution across the defined categories.

Section 6 describes the data governance mechanisms applied specifically during the assessment process. It documents the automated sensitive data detection system deployed to identify articles requiring human review prior to admission into the corpus, the quarantine and human review workflow through which flagged articles were processed, and the traceability mechanisms that ensure every assessment decision is recorded and auditable. This section does not reproduce the governance framework established in D7.1 but documents its operational implementation within the specific context of the assessment pipeline.

Section 7 presents the conclusions of the deliverable and a set of recommendations directed at the subsequent tasks and deliverables of Work Package 7. The conclusions synthesise the key findings of the assessment process and their implications for the design of the preprocessing pipeline, the structure of the storage infrastructure, and the LLM training activities conducted in Work Package 2. The recommendations are formulated as actionable guidance for the teams responsible for D7.3, D7.4, and D7.5.

The document concludes with three annexes. Annex A provides the assessment scorecard template applied during the evaluation process, enabling the reader to understand the precise instrument through which individual articles were scored across the assessment dimensions described in Section 3. Annex B lists the MeSH descriptors applied during the categorisation process documented in Section 5. Annex C presents the corpus statistics tables that underpin the quantitative findings reported in Sections 4 and 5.

2 Overview of the Data Collection Process

The data assessment and categorisation activities documented in this deliverable were conducted on a corpus of biomedical and clinical scientific literature assembled by Bosonit S.L. during the first phase of Work Package 7. This section provides a factual account of how that corpus was built: the repositories from which literature was sourced, the methodology applied to identify, retrieve, and store eligible articles, and the general scope and volume of the resulting collection. The legal eligibility criteria that governed source selection including licensing requirements, rights clearance procedures, and exclusion rules based on content sensitivity are established in full in D7.1 and are not reproduced here. This section treats those criteria as operative constraints and describes only the collection process as it was implemented within those boundaries.

2.1 Selected repositories

The selection of data sources for the AIDESL corpus was the result of a systematic evaluation of the principal repositories of open-access biomedical and clinical scientific literature available at the international level. The evaluation considered four primary factors: thematic alignment with the project's target application domains, the availability of a public programmatic interface enabling systematic and reproducible retrieval of articles and associated metadata, the licensing model governing the deposited content, and the volume and recency of the literature available within the relevant subject areas. Following this evaluation, three repositories were selected as the primary sources for the AIDESL corpus.

medRxiv is a preprint server dedicated to the health sciences, operated by Cold Spring Harbor Laboratory in collaboration with the British Medical Journal and Yale University. It constitutes the primary source of literature for the AIDESL corpus by virtue of its direct and comprehensive coverage of the project's three target domains: medical device development, pharmaceutical evaluation, and clinical research. medRxiv provides a public API that enables the programmatic retrieval of article metadata and full-text PDFs through Digital Object Identifier (DOI)-based queries, which aligns directly with the ingestion mechanism implemented in the WP7 data pipeline. The repository operates under an open-access model in which deposited articles are made available under Creative Commons licences, predominantly CC BY and CC BY-NC, with the specific licence of each article retrievable through the API response and recorded as part of the article's metadata at the point of ingestion.

bioRxiv is the life sciences counterpart of medRxiv, also operated by Cold Spring Harbor Laboratory. Although its primary focus is biological sciences rather than clinical medicine, bioRxiv provides relevant literature for the AIDESL corpus in areas adjacent to the project's target domains, including molecular biology, genomics, and translational research. Its API shares the same structure and query interface as medRxiv, which enabled the reuse of the ingestion pipeline developed for the primary source with minimal adaptation. Content licensing conditions are equivalent to those of medRxiv, with articles predominantly available under Creative Commons licences.

arXiv is a multidisciplinary preprint repository operated by Cornell University, covering physics, mathematics, computer science, quantitative biology, and related fields. Its inclusion in the AIDESL source portfolio responds to the need to capture the rapidly growing body of methodological literature at the intersection of artificial intelligence, natural language processing, and biomedical research a domain of particular relevance to the technical development activities of Work Packages 2 and 3. arXiv provides a public API and bulk data access mechanisms, and the majority of its content is available under open licences compatible with the project's use requirements as established in D7.1.

The three repositories were selected as the exclusive primary sources for the AIDESL corpus. Subscription-based databases, including PubMed/MEDLINE and Europe PMC, were evaluated during the source selection process but were not incorporated as primary ingestion sources, given the licensing and contractual restrictions that would have applied to their systematic use for LLM training purposes under the compliance framework established in D7.1. Literature indexed in those databases that is simultaneously deposited in one of the three selected preprint repositories remains accessible to the project through those open-access channels.

2.2 Collection methodology and scope

The construction of the AIDESL corpus proceeded in two distinct phases. The first phase, conducted during the initial months of the project following its launch in October 2024, consisted of a structured bulk collection exercise targeting the three selected repositories. During this phase, the ingestion pipeline developed under Task 7.3 was deployed systematically to retrieve literature across the project's three target domains, applying the thematic and licensing filters established in D7.1 to ensure that only eligible content was admitted into the corpus. This initial bulk collection established the foundational corpus upon which the assessment activities documented in subsequent sections of this deliverable were primarily conducted.

The second phase, which continued through the remainder of the period covered by this deliverable, consisted of targeted supplementary ingestion operations conducted to address specific gaps identified during the assessment process, to incorporate literature published after the initial bulk collection window, and to validate the ingestion pipeline against newly available content. These supplementary operations did not alter the source portfolio or the eligibility criteria but refined the retrieval parameters in response to the empirical findings reported in Section 4.

Article retrieval was conducted programmatically through the public APIs of the three selected repositories, using DOI-based queries as the primary retrieval mechanism. For each article, the pipeline retrieved both the full-text PDF and the associated structured metadata, including title, authors, publication date, repository of origin, and licence type. The DOI was extracted directly from the content of the retrieved PDF where not returned by the API, using a pattern-matching procedure applied to the initial section of the document prior to any cleaning or transformation step. All retrieved articles and their associated metadata were stored in the raw data layer of the pipeline, from which they proceeded to the assessment process described in Section 3.

With regard to temporal scope, the corpus was bounded to literature published from January 2019 onwards. This lower boundary was established on the basis of two considerations. First, the period from 2019 onwards witnessed a substantial acceleration in the volume of preprint deposition across the three selected repositories, particularly in the clinical and biomedical domains, driven in part by the expansion of open-access publishing practices and the surge in research output associated with the COVID-19 pandemic. Second, literature published prior to 2019 was considered less representative of the methodological standards and reporting conventions that characterise the current state of the art in systematic literature review practice, and therefore of lower value for the training of models intended to support contemporary SLR workflows. No upper temporal boundary was applied; the corpus includes articles available at the time of each ingestion operation up to the submission date of this deliverable.

With regard to thematic scope, retrieval queries were designed to target literature relevant to the project's three primary application domains medical device development, pharmaceutical evaluation, and clinical research as well as the methodological domain of AI and NLP applied to biomedical

literature processing. Query terms were defined in consultation with the domain experts within the consortium and refined iteratively based on the precision and recall characteristics observed during the initial bulk collection phase.

2.3 Volume and general characteristics of the corpus

The corpus assembled through the collection process described in Section 2.2 comprises several hundred articles drawn from the three selected repositories. The distribution across repositories reflects the differential coverage of each source relative to the project's target domains: medRxiv accounts for the largest share of the corpus, consistent with its position as the primary source for clinical and health sciences literature; bioRxiv contributes a secondary but significant share of content in adjacent biological and translational research areas; and arXiv provides a more targeted collection of methodological and AI-focused literature relevant to the technical development activities of the project.

In terms of document types, the corpus encompasses primary research articles, systematic reviews, meta-analyses, clinical trial reports, and observational studies. Articles presenting methodological contributions particularly those introducing or evaluating NLP and AI techniques for biomedical literature processing are also represented, sourced primarily from arXiv. Editorial content, correspondence, and opinion pieces were excluded from the corpus at the point of ingestion in accordance with the document type eligibility criteria established in D7.1.

All articles in the corpus are written in English, reflecting both the retrieval parameters applied during ingestion and the operational characteristics of the processing pipeline, which was designed and validated for English-language content. The temporal distribution of the corpus is concentrated in the period from 2020 onwards, with the highest density of articles corresponding to the years 2021 through 2025, consistent with the growth trajectory of preprint deposition in the biomedical domain over this period.

A more detailed quantitative characterisation of the corpus, disaggregated by repository, domain, document type, and temporal distribution, is presented in Annex C and discussed in the context of the assessment findings in Section 4.

3 Assessment Methodology

The assessment of the literature collected under Task 7.3 required the development of a dedicated evaluation framework capable of determining, for each article in the corpus, whether and to what degree that article constitutes a valuable contribution to the AIDESL training dataset. This determination cannot be reduced to a binary eligibility judgement of the kind applied during the collection phase, where the operative criteria were primarily legal and structural in nature, as established in D7.1. The assessment process addressed a qualitatively different set of questions: not whether an article may lawfully be included in the corpus, but whether it should be included given its scientific content, its thematic alignment with the project's objectives, and its practical suitability for the specific technical task of LLM training on biomedical literature.

The assessment framework described in this section was developed entirely by Bosonit S.L. as part of its leadership of Work Package 7. It was not inherited from an existing standard or external specification but was constructed iteratively in response to the characteristics of the literature encountered during the initial bulk collection phase and the technical requirements communicated by the teams responsible for

LLM development and training within Work Package 2. The framework represents a methodological contribution of the project in its own right, and its design reflects a number of deliberate choices that are explained and justified in the subsections that follow.

The framework is structured around three principal analytical dimensions, each of which captures a distinct and non-redundant aspect of an article's value to the project. The first dimension, relevance, addresses the degree to which the content of an article pertains to the project's target application domains and use cases. The second dimension, scientific quality, addresses the methodological rigour, reporting standards, and evidential value of the article as a scientific document. The third dimension, suitability for LLM training, addresses the structural, linguistic, and formatting characteristics of the article that determine how effectively it can be processed, segmented, and used as training material by the pipeline developed under WP7. These three dimensions are assessed through a structured scoring approach applied consistently across the corpus, as described in Section 3.2, Section 3.3, and Section 3.4 respectively.

An important characteristic of the assessment framework is that its three dimensions are treated as independent axes of evaluation rather than as components of a single composite score. This design choice was made deliberately to preserve the analytical granularity of the assessment and to avoid situations in which a high score on one dimension masks a disqualifying deficiency on another. An article that is highly relevant to the project's target domains but is structurally unsuitable for automated processing, for instance, would receive a high relevance score and a low suitability score, and this profile would be treated differently by the downstream pipeline than an article with moderate scores across all three dimensions. The practical implications of each scoring profile for the downstream use of the assessed articles are described in Section 4.

The assessment framework was applied in two modalities across the corpus. The primary modality consisted of automated scoring, in which structured heuristics and natural language processing tools were applied programmatically to each article to produce initial scores across the three dimensions. The secondary modality consisted of human review, applied to a stratified sample of the corpus to validate the automated scores and to handle cases flagged by the governance layer as requiring expert judgement. The interaction between automated scoring and human review, and the governance mechanisms through which that interaction was managed, are described in Section 6.

Section 3.5 describes the overall scoring and ranking approach, including the decision rules applied to translate dimension scores into operational classifications that determine how each article is treated by the downstream pipeline.

3.1 Assessment framework and dimensions

The design of the assessment framework began with a requirements-gathering exercise conducted by Bosonit S.L. during the early months of the project. This exercise had two primary inputs. The first was a systematic review of the characteristics of the literature encountered during the initial bulk ingestion operations, which revealed significant heterogeneity in the thematic focus, methodological approach, document structure, and linguistic register of the articles retrieved from the three selected repositories. The second input was a set of technical requirements derived from the specifications of the LLM training and inference pipeline under development within Work Package 2, which imposed concrete constraints on the minimum content density, structural predictability, and semantic specificity required of training documents for them to contribute meaningfully to model performance.

From these two inputs, Bosonit S.L. derived the three-dimensional framework described in this section. The following paragraphs describe the rationale and conceptual basis of each dimension before the detailed operationalisation of each is presented in Sections 3.2, 3.3, and 3.4.

The relevance dimension addresses the fundamental question of whether the subject matter of an article contributes to the knowledge base that the AIDESL platform is intended to encode and exploit. AIDESL targets three primary application domains: medical device development, pharmaceutical evaluation, and clinical research. Each of these domains encompasses a wide range of specific topics, methodologies, and research questions, and the boundary between in-scope and out-of-scope content is not always immediately apparent from titles or abstracts alone. A pharmacokinetic study of a novel compound, for instance, is clearly within scope for the pharmaceutical evaluation domain; a purely computational study of protein folding algorithms may or may not be relevant depending on whether its results have direct implications for drug development or clinical decision-making. The relevance dimension was designed to capture this spectrum of thematic alignment in a graded rather than binary manner, enabling the assessment process to distinguish between articles that are centrally relevant, peripherally relevant, and effectively out of scope, and to treat each category differently in the downstream pipeline.

The scientific quality dimension addresses the epistemic value of an article as a contribution to the biomedical knowledge base. This dimension is particularly important in the context of LLM training because models trained on low-quality scientific content risk encoding imprecise, inconsistent, or methodologically flawed information that could compromise the reliability of the extraction outputs they subsequently produce. The preprint repositories selected as sources for the AIDESL corpus do not impose the same quality controls as peer-reviewed journals: articles deposited in medRxiv, bioRxiv, and arXiv have not undergone editorial peer review at the point of deposition, and the range of methodological rigour represented in these repositories is consequently wide. The scientific quality dimension was designed to function as a proxy for the peer review process within the assessment framework, applying a structured set of quality indicators to each article to estimate its methodological robustness and the reliability of its reported findings.

The suitability for LLM training dimension addresses a set of characteristics that are orthogonal to both relevance and scientific quality but are equally determinative of an article's practical value to the project. An article may be highly relevant and scientifically rigorous but nonetheless unsuitable for LLM training if, for instance, its content is dominated by figures and tables with minimal prose, its text layer is corrupted or incomplete due to PDF rendering issues, its language is highly specialised to a degree that falls outside the distributional range of the base models used in the project, or its structure deviates significantly from the canonical sections of a biomedical research article in ways that impair the segmentation and chunking operations applied by the preprocessing pipeline. The suitability dimension was designed to capture these technical and structural characteristics systematically, ensuring that the assessment process produces actionable guidance not only on which articles to include in the corpus but on how each included article should be handled by the downstream pipeline.

The three dimensions were formalised into a structured assessment instrument, referred to throughout this document as the assessment scorecard, whose full specification is provided in Annex A. The scorecard defines, for each dimension, a set of indicators, each with an associated scoring scale and decision rule. Indicators were selected on the basis of two criteria: first, that they capture information that is genuinely predictive of the article's value to the project along the relevant dimension; and second, that they can be evaluated consistently and reproducibly, whether through automated scoring or human review. Indicators that would require deep domain expertise to evaluate consistently and could not be

reliably approximated through automated means were excluded from the scorecard or relegated to the human review modality.

The scorecard was validated prior to full deployment through a calibration exercise in which a sample of articles drawn from each of the three repositories was independently assessed by two members of the Bosonit S.L. technical team using the draft scorecard. The results of this calibration exercise were used to identify indicators where inter-rater agreement was insufficient, to refine the scoring scales and decision rules associated with those indicators, and to establish the automated scoring heuristics that were subsequently applied to the full corpus. The calibration exercise also provided the first empirical data on the distribution of scores across the three dimensions, which informed the design of the scoring and ranking approach described in Section 3.5.

A final design principle of the assessment framework warrants explicit mention. The framework was conceived from the outset as a living instrument rather than a fixed specification. As the assessment process progressed and the characteristics of the corpus became better understood, refinements were introduced to the scoring scales, decision rules, and automated heuristics in a controlled manner, with each revision documented and version-controlled to ensure the traceability of all assessment decisions. This approach reflects the inherently iterative nature of corpus construction for LLM training and is consistent with the adaptive methodology described in Section 1.1. The version of the assessment scorecard applied to the bulk of the corpus, and the one whose outputs underpin the findings reported in Section 4, is the version provided in Annex A.

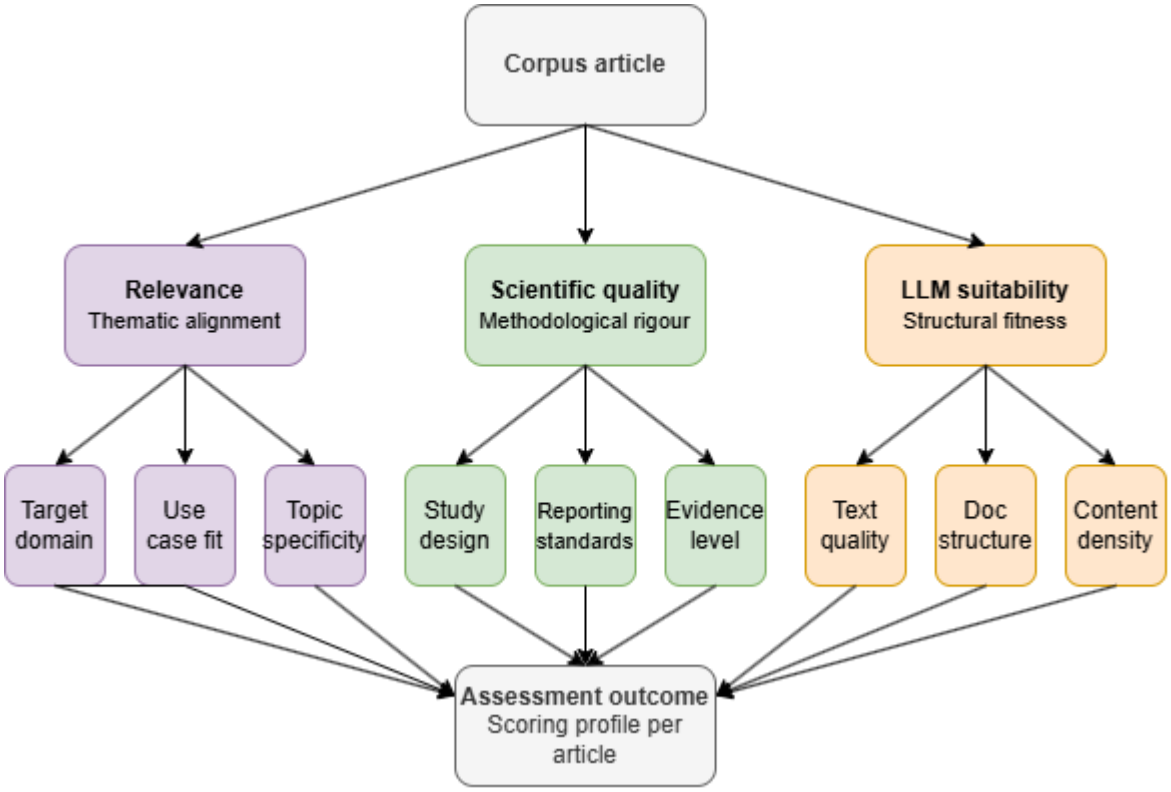

Figure 1. Three-dimensional assessment framework applied to the AIDESL literature corpus

3.2 Relevance assessment

The relevance dimension of the assessment framework addresses the degree to which the substantive content of an article contributes to the thematic knowledge base that the AIDESL platform is designed to encode and exploit. Relevance, as operationalised in this framework, is not a binary property but a graded one, reflecting the observation that the relationship between a given article and the project's target domains can range from direct and central to peripheral and tangential, with a wide intermediate spectrum that requires careful evaluation.

The AIDESL project is oriented towards three primary application domains: medical device development, clinical research, and pharmaceutical evaluation. Each of these domains has a distinct knowledge profile in terms of the types of studies it encompasses, the outcome measures it typically reports, the regulatory frameworks it operates within, and the methodological conventions it applies. An article that speaks directly to one of these domains, addressing its characteristic research questions and reporting its findings in the terminology and format conventions of that domain, is of high relevance to the project. An article that addresses a topic adjacent to one of the domains but does not engage with its characteristic concerns, or that addresses a methodological question that spans all three domains without being anchored in any of them, occupies a lower relevance tier. An article whose content, however scientifically valuable in its own right, falls outside the thematic scope of any of the three target domains is assigned the lowest relevance tier and is excluded from the core training corpus, though it may be retained in a supplementary partition for methodological purposes.

The relevance assessment was operationalised through three indicators, each evaluated independently and recorded in the assessment scorecard.

The first indicator is target domain alignment. This indicator captures the degree to which the article's primary subject matter pertains to one or more of the three AIDESL target domains. Assessment of this indicator was conducted primarily through automated classification, using a keyword-based matching procedure applied to the article's title, abstract, and section headings. The keyword vocabulary was developed iteratively by the Bosonit S.L. technical team during the calibration phase, drawing on domain terminology derived from Medical Subject Headings (MeSH) controlled vocabulary in the relevant thematic branches and refined through manual inspection of a representative sample of articles from each repository. Articles for which the automated classifier returned ambiguous results were escalated to human review.

The second indicator is use case fit. Where the target domain alignment indicator assesses the broad thematic territory of an article, the use case fit indicator assesses the specific practical relevance of the article's content to the extraction tasks and downstream applications that the AIDESL platform is designed to support. The project's three demonstration scenarios, as defined in Work Package 5, impose specific requirements on the type of information that the trained models must be capable of extracting: clinical efficacy and safety outcomes in the context of medical device evaluation, pharmacokinetic and pharmacodynamic parameters in the context of pharmaceutical studies, and patient population characteristics, intervention descriptions, and comparative outcomes in the context of clinical research. The use case fit indicator scores each article according to whether its content includes information of the type targeted by one or more of these extraction tasks, irrespective of whether that content is framed in terms explicitly linked to the AIDESL demonstration scenarios.

The third indicator is topic specificity. This indicator addresses the level of thematic precision of the article's contribution to its primary domain. A highly specific article is one whose subject matter is narrowly defined and whose findings are directly actionable within the relevant domain; a low-specificity

article is one whose subject matter is broad, whose findings are generic across multiple domains, or whose content is predominantly methodological or theoretical without direct domain application. Topic specificity was assessed primarily through automated heuristics based on the density of domain-specific terminology in the article's body text, calibrated against the calibration sample to account for differences in terminological density across the three repositories.

Each of the three indicators was scored on a three-point scale: high, moderate, or low. The composite relevance profile of each article was determined by the combination of its scores across the three indicators, with no single indicator acting as a veto in isolation. Articles presenting a combination of high target domain alignment, high use case fit, and high or moderate topic specificity were assigned to the core relevance tier and prioritised for inclusion in the primary training corpus. Articles with mixed profiles, including those with high domain alignment but low use case fit or vice versa, were assigned to the secondary relevance tier and included in the training corpus with appropriate metadata flags that enable the downstream pipeline to weight them appropriately during training. Articles scoring low on all three indicators were excluded from the training corpus and documented in the exclusion log maintained as part of the traceability system described in Section 6.

The relevance assessment was the first of the three dimensions to be applied in the sequential scoring pipeline, functioning as an initial filter that determined whether an article proceeded to quality and suitability assessment. Articles assigned to the exclusion tier on relevance grounds were not scored on the remaining two dimensions, which both reduced the computational cost of the assessment process and ensured that quality and suitability scores were produced only for articles with a genuine prospect of inclusion in the corpus.

3.3 Quality assessment

The scientific quality dimension of the assessment framework addresses the methodological rigour, reporting standards, and evidential value of each article as a scientific document. This dimension is of particular importance in the context of LLM training because the quality of the information encoded in a language model during training is directly conditioned by the quality of the documents from which that information is drawn. A model trained on a corpus that includes a significant proportion of methodologically weak, poorly reported, or evidentially unreliable literature risks encoding imprecise or inconsistent representations of biomedical knowledge that would compromise the reliability and accuracy of the extraction outputs it subsequently produces. The scientific quality dimension was therefore designed to function as a systematic proxy for the peer review and editorial quality control processes that the preprint repositories selected as sources for the AIDESL corpus do not themselves provide.

This design choice warrants explicit justification. The decision to source literature from preprint repositories rather than peer-reviewed journals was motivated by the legal and licensing considerations documented in D7.1, which identified open-access preprint repositories as the most reliably compliant source of biomedical literature for LLM training purposes. The consequence of this decision, however, is that the articles in the corpus span a significantly wider range of methodological quality than would be the case in a corpus sourced exclusively from high-impact, peer-reviewed journals. Articles deposited in medRxiv, bioRxiv, and arXiv are made available without prior editorial peer review, and the range of methodological approaches, reporting practices, and evidential standards represented across these repositories is correspondingly heterogeneous. The scientific quality assessment dimension was developed precisely to address this heterogeneity in a systematic and reproducible manner, enabling the project to benefit from the legal accessibility of preprint literature while mitigating the quality risks that this source selection entails.

The scientific quality assessment was operationalised through three indicators, each designed to capture a distinct and non-redundant aspect of an article's methodological and evidential standing.

The first indicator is study design robustness. This indicator assesses the appropriateness and rigour of the methodological approach adopted in the article relative to the research question it addresses. In the context of biomedical research, study design is a primary determinant of the reliability of reported findings: a well-designed randomised controlled trial provides a higher level of evidence for a causal claim than an observational cohort study, which in turn provides stronger evidence than a case series or an expert opinion piece. The assessment framework does not apply a simple hierarchy of study designs, however, since different designs are appropriate for different types of research questions, and a well-executed observational study addressing a question for which a randomised trial would be impractical or unethical represents a higher quality contribution than a poorly executed trial addressing the same question.

Study design robustness was assessed using a two-stage procedure. In the first stage, the automated scoring pipeline classified each article by study type using a combination of keyword matching against the article's methods section and a structured search for reporting guideline adherence markers, specifically the presence of terms associated with established reporting standards such as CONSORT for randomised trials, STROBE for observational studies, PRISMA for systematic reviews, and ARRIVE for preclinical studies. In the second stage, articles for which the automated classification returned an ambiguous or low-confidence result were escalated to human review, where a member of the Bosonit S.L. technical team applied the indicator scoring criteria directly to the article's methods section. The automated classification achieved satisfactory agreement with the human review scores across the calibration sample, with discrepancies concentrated in articles reporting mixed-methods or hybrid study designs that did not map cleanly to any single reporting standard category.

The second indicator is reporting completeness and transparency. This indicator addresses the degree to which an article provides sufficient information about its methods, data, results, and limitations to allow an informed reader to assess the validity of its findings and, where relevant, to replicate its analysis. Transparent and complete reporting is a prerequisite for the reliable extraction of structured information from scientific literature: an article that omits key methodological details, presents results without adequate statistical context, or fails to acknowledge limitations provides an incomplete and potentially misleading representation of the evidence that would be encoded as such in any model trained on it.

Reporting completeness was assessed through a structured checklist applied to each article, covering the presence and adequacy of descriptions of the study population or sample, the intervention or exposure under investigation, the outcome measures and their operationalisation, the statistical methods applied, and the key limitations acknowledged by the authors. The checklist was derived from the core elements common to the major reporting guidelines referenced above, adapted to apply across multiple study types without requiring type-specific scoring criteria for each. Automated scoring of this indicator was based on a combination of section presence detection, which verified that the article contained identifiable methods, results, and discussion sections, and keyword density analysis applied to those sections to estimate the richness of methodological and statistical reporting. Human review was applied to articles for which the automated scores fell in the intermediate range, where the distinction between adequate and inadequate reporting was most consequential for corpus inclusion decisions.

The third indicator is evidence level and reliability. This indicator addresses the position of the article's contribution within the hierarchy of evidence that governs biomedical research, and the degree to which the reliability of its reported findings is supported by the strength of its design and the robustness of its analysis. Evidence level in biomedical research is a well-established concept, with systematic reviews

and meta-analyses of randomised controlled trials occupying the apex of the hierarchy and expert opinion or case reports at its base. For the purposes of the AIDESL corpus, articles at higher levels of the evidence hierarchy are assigned higher scores on this indicator not because lower-level evidence is without value but because higher-level evidence articles provide richer, more structured, and more semantically precise representations of biomedical knowledge that are of greater value for LLM training on the extraction tasks targeted by the project.

Evidence level was assessed through a combination of study type classification, which provided a first approximation of evidence level based on design, and a secondary assessment of the quality and consistency of the statistical evidence presented in the article's results section. The latter assessment focused on the presence of measures of uncertainty such as confidence intervals, p-values, and effect sizes, the consistency of reported findings across subgroup analyses where applicable, and the authors' own characterisation of the strength of the evidence in the discussion and conclusions sections. Articles reporting primary findings without adequate statistical context, or whose conclusions overstated the strength of the evidence relative to the design and sample size, were penalised on this indicator irrespective of their study type classification.

As with the relevance dimension, each of the three scientific quality indicators was scored on a three-point scale and the composite quality profile of each article was determined by the combination of its scores across the three indicators. Articles presenting strong profiles across all three indicators were assigned to the high-quality tier of the corpus and treated as priority training documents for the model development activities of Work Package 2. Articles with moderate composite profiles were included in the corpus with quality tier metadata flags, enabling the training pipeline to apply differential weighting where appropriate. Articles with weak profiles on two or more indicators were excluded from the training corpus and documented in the exclusion log, with the specific indicators responsible for the exclusion recorded to support the retrospective analysis of corpus quality reported in Section 4.

One important methodological consideration warrants explicit acknowledgement. The scientific quality assessment described in this section is necessarily approximate in a number of respects. The automated scoring heuristics applied to study design robustness and reporting completeness capture observable proxies for quality rather than quality itself: the presence of reporting guideline markers indicates that an author was aware of established standards and attempted to comply with them, but does not guarantee that the reported study was in fact conducted with the rigour those standards prescribe. Similarly, the evidence level assessment based on study type classification provides a reasonable approximation of evidential standing in the majority of cases but may misclassify articles whose study design is atypical or whose reporting conventions deviate from the norms of their study type category. These limitations are acknowledged in the assessment scorecard provided in Annex A and are reflected in the uncertainty estimates associated with the quality tier assignments reported in Section 4. The human review procedures described in Section 6 were designed in part to mitigate these limitations by ensuring that articles in the intermediate scoring range, where the approximations inherent in automated scoring are most consequential, are subject to expert assessment before their corpus inclusion status is finalised.

3.4 Suitability for LLM training

The suitability for LLM training dimension of the assessment framework addresses a set of characteristics that are orthogonal to both relevance and scientific quality but are equally determinative of an article's practical value to the project. An article may score highly on both preceding dimensions and yet be unsuitable as training material if its physical and structural properties prevent the data pipeline from processing it correctly, if its textual content is too sparse or too specialised to contribute meaningfully to model learning, or if its formatting conventions deviate from the structural patterns that

the extraction pipeline expects. The suitability dimension was designed to surface these technical and structural constraints systematically, ensuring that the assessment process produces actionable guidance not only on which articles to include in the corpus but on how each included article should be handled by the downstream processing pipeline.

The importance of this dimension reflects a characteristic of LLM training that distinguishes it from many other uses of scientific literature. When a researcher reads a scientific article, they apply a rich set of cognitive strategies to extract meaning from imperfect textual signals: they infer content from figures that they can see, they reconstruct arguments across poorly typeset passages, they interpret abbreviations from context, and they discount sections that are clearly ancillary to the main contribution. An LLM training pipeline cannot apply these strategies. It operates on the text layer of the article as extracted from the PDF, and the quality, completeness, and structural coherence of that text layer determine the quality of the training signal that the article contributes. An article whose text layer is substantially degraded, fragmented, or dominated by non-prose content provides a weaker training signal than its scientific content would suggest, and in extreme cases may introduce noise into the training process that is actively detrimental to model performance.

The suitability assessment was operationalised through three indicators, each targeting a distinct aspect of an article's fitness for automated processing and LLM training.

The first indicator is text extraction quality. This indicator assesses the degree to which the full-text content of the article can be reliably extracted from its PDF representation and rendered as clean, coherent, machine-readable prose. PDF extraction of scientific literature is a technically demanding operation due to the structural complexity of scientific documents: multi-column layouts, mathematical formulae, embedded figures and tables, footnotes, and running headers all introduce potential failure modes that can result in garbled, incomplete, or incorrectly ordered text output. The severity of these failure modes varies significantly across articles depending on the PDF generation method, the typesetting software used, and the complexity of the document's visual layout.

Text extraction quality was assessed by applying the extraction procedure implemented in the WP7 pipeline to each article and evaluating the quality of the resulting text output against a set of diagnostic criteria. These criteria included the proportion of the article's estimated word count successfully recovered in the extracted text, the coherence of the extracted text as measured by the absence of anomalous character sequences indicative of encoding errors, the correct ordering of extracted text segments relative to the logical reading order of the document, and the absence of significant contamination of the main body text by header, footer, or reference list content. Articles for which the extraction procedure yielded text meeting all diagnostic criteria were assigned a high score on this indicator. Articles for which the extraction procedure yielded text with moderate deficiencies, such as occasional encoding anomalies or minor ordering errors confined to non-critical sections, were assigned a moderate score and flagged for preprocessing intervention at the pipeline stage described in D7.4. Articles for which the extraction procedure yielded substantially degraded or incomplete text, representing less than a defined minimum proportion of the estimated document content, were assigned a low score and excluded from the training corpus on suitability grounds irrespective of their scores on the relevance and quality dimensions.

The second indicator is document structure coherence. This indicator assesses the degree to which the article conforms to the canonical sectional structure of a biomedical research publication, specifically the presence and identifiability of the standard sections that define the organisation of a scientific article in the domains covered by the AIDESL corpus. These sections, which correspond broadly to the IMRAD structure widely adopted in biomedical publishing, encompass the abstract, introduction, methods,

results, and discussion, together with ancillary sections such as conclusions, funding statements, and conflict of interest disclosures. The identifiability of these sections is of operational significance because the preprocessing pipeline implemented under WP7 applies section-specific processing strategies that depend on the correct identification of section boundaries: the abstract is treated as a high-density summary from which metadata and classification signals are derived, the methods section is the primary source of study design and procedural information, and the results section is the primary source of outcome and finding data.

Document structure coherence was assessed through a section detection procedure applied to the extracted text of each article, which identified section headings and mapped them to the canonical section taxonomy using a combination of exact matching and fuzzy matching against a controlled vocabulary of section heading variants. Articles for which all five core IMRAD sections were reliably identified were assigned a high score on this indicator. Articles for which three or four core sections were identified, with the missing sections attributable to legitimate reporting conventions of the article's study type rather than to structural deficiencies in the document, were assigned a moderate score. Articles for which fewer than three core sections were identifiable, or for which the section detection procedure returned inconsistent or contradictory results, were assigned a low score and flagged for human review to determine whether the structural deficiency was attributable to the document itself or to a failure of the detection procedure.

The third indicator is content density and prose quality. This indicator addresses the proportion of an article's text that consists of substantive, information-bearing prose as opposed to non-prose content that contributes little to the training signal. Non-prose content in scientific articles encompasses several categories that are particularly prevalent in biomedical literature: numerical tables presenting raw data, figure legends describing visual content that is not present in the text layer, reference lists, statistical notation and formulae, and boilerplate text such as ethics statements, author contribution statements, and data availability declarations. While none of these categories is entirely without value as training material, their informational density per token is substantially lower than that of the article's main body prose, and a high proportion of non-prose content reduces the effective informational yield of an article relative to its raw word count.

Content density was assessed through a combination of prose detection heuristics applied to the extracted text of each article. These heuristics identified and quantified the proportion of the extracted text attributable to reference lists, based on the characteristic formatting of bibliographic citations; to tabular content, based on the presence of tab-separated or structured numerical data patterns; and to figure legends and captions, based on the presence of figure and table reference markers followed by descriptive text. The residual proportion of the text, after exclusion of these categories, was treated as an approximation of the substantive prose content of the article and used as the basis for the content density score. Articles with a high substantive prose proportion were assigned a high score on this indicator. Articles with a moderate proportion, typically those with a high density of tables or supplementary appendices, were assigned a moderate score and flagged to inform the chunking strategy applied during preprocessing. Articles in which the substantive prose proportion fell below a defined minimum threshold were assigned a low score, reflecting the assessment that their contribution to the training signal would be insufficient to justify the processing overhead of their inclusion.

Beyond the quantitative assessment of content density, the indicator also incorporated a qualitative assessment of prose quality in a sample of articles subject to human review. Prose quality in this context refers to the linguistic characteristics of the article's body text that determine its utility as training material, including sentence structure complexity, terminological consistency, and the absence of significant linguistic artefacts attributable to non-native English writing or to automated translation.

While these characteristics cannot be reliably assessed through automated heuristics alone at scale, their potential impact on training quality is significant, and the human review sample was designed to provide a statistically valid estimate of the prevalence of prose quality issues across the corpus as a whole. The findings of this estimation are reported in Section 4.

The composite suitability profile of each article was determined by the combination of its scores across the three indicators, applying the same three-tier classification used for the relevance and quality dimensions. Articles with high scores across all three suitability indicators were admitted to the training corpus without any pipeline-specific processing flags. Articles with moderate composite suitability profiles were admitted with pipeline flags that activated targeted preprocessing interventions at the appropriate stages of the WP7 processing pipeline, as detailed in D7.4. Articles with low scores on the text extraction quality indicator were excluded from the training corpus irrespective of their scores on the remaining suitability indicators, given that the reliability of downstream assessment and processing depends fundamentally on the availability of a high-quality text representation. Articles with low scores on document structure coherence or content density in isolation were subject to a case-by-case determination during human review, taking into account the article's profiles on the relevance and quality dimensions and the potential value of targeted preprocessing interventions in recovering a usable training signal.

3.5 Scoring and ranking approach

The three assessment dimensions described in Sections 3.2, 3.3, and 3.4 each produce an independent dimensional profile for every article in the corpus, expressed as a three-point categorical score of high, moderate, or low across each of the three indicators within the dimension. The scoring and ranking approach described in this section defines how these dimensional profiles are combined into a single operational classification that determines the treatment of each article by the downstream pipeline. The approach was designed to preserve the analytical granularity of the three-dimensional assessment while producing unambiguous and consistently applicable decision rules that the pipeline can execute without requiring case-by-case human judgement for the majority of articles.

The scoring system operates in two stages. The first stage aggregates the three indicator scores within each dimension into a dimensional tier assignment. The second stage combines the three dimensional tier assignments into an overall corpus classification that maps directly to a set of operational pipeline instructions.

Dimensional tier assignment

Within each of the three assessment dimensions, the three indicator scores are aggregated into a single dimensional tier using a rule-based procedure rather than a simple average. The rationale for using rules rather than averages is that within each dimension, certain indicators function as stronger constraints than others, and a simple average would mask disqualifying deficiencies on those indicators. The specific aggregation rules applied within each dimension reflect the relative criticality of its constituent indicators as determined during the calibration phase.

For the relevance dimension, an article is assigned to the high relevance tier if it scores high on target domain alignment and high or moderate on at least one of the remaining two indicators. It is assigned to the moderate relevance tier if it scores high on target domain alignment but low on both remaining indicators, or if it scores moderate on target domain alignment irrespective of its scores on the remaining indicators. It is assigned to the low relevance tier if it scores low on target domain alignment, which acts

as a necessary but not sufficient condition for corpus inclusion. This rule reflects the primacy of domain alignment as the fundamental relevance criterion: an article that does not engage with any of the three target domains cannot contribute meaningfully to the AIDESL training corpus regardless of its use case fit or topic specificity scores.

For the scientific quality dimension, an article is assigned to the high quality tier if it scores high or moderate on study design robustness and high or moderate on reporting completeness, irrespective of its evidence level score. It is assigned to the moderate quality tier if it scores high on either study design robustness or reporting completeness but low on the other, or if it scores moderate on both. It is assigned to the low quality tier if it scores low on both study design robustness and reporting completeness simultaneously, which is treated as indicative of a fundamental methodological deficiency that cannot be compensated by a high evidence level score. This rule reflects the assessment that methodological transparency and reporting completeness are the most consequential quality characteristics for training data purposes, since they determine the reliability of the information that the article contributes to the corpus independently of the formal evidence level of its study design.

For the suitability dimension, the text extraction quality indicator acts as a hard constraint: any article scoring low on text extraction quality is assigned to the low suitability tier irrespective of its scores on the remaining two indicators. Subject to this constraint, an article is assigned to the high suitability tier if it scores high on both document structure coherence and content density. It is assigned to the moderate suitability tier if it scores high on one and moderate or low on the other, or moderate on both. It is assigned to the low suitability tier, in addition to the text extraction quality constraint case, if it scores low on both document structure coherence and content density. This rule reflects the operational dependency of the downstream pipeline on text extraction quality as a prerequisite for all subsequent processing steps.

Overall corpus classification

The three dimensional tier assignments are combined into one of four overall corpus classifications, each of which maps to a specific set of pipeline instructions. The four classifications are designated as follows.

Articles classified as Core corpus are those that achieve high or moderate tier assignments across all three dimensions, with at least two high tier assignments among the three. These articles are treated as primary training documents and are processed through the full pipeline without any dimension-specific flags. They constitute the highest-priority segment of the training corpus and are the articles against which the performance of the trained models is primarily evaluated in the validation activities described in D7.6.

Articles classified as Flagged corpus are those that achieve a combination of dimensional tier assignments that includes at least one moderate and no more than one low, provided that the low tier assignment does not occur on the suitability dimension's text extraction quality indicator, which would trigger exclusion. These articles are included in the training corpus but are processed with one or more pipeline flags that activate targeted interventions at the relevant processing stages. The specific flags applied to each article are recorded as metadata fields in the article's Silver layer object and are consumed by the preprocessing pipeline documented in D7.4. Flagged corpus articles are weighted differently from Core corpus articles during model training, with the specific weighting scheme determined by the WP2 team responsible for LLM development based on the distribution of flags across the corpus.

Articles classified as Supplementary corpus are those that achieve a high or moderate tier assignment on the relevance dimension but a low tier assignment on either the quality or suitability dimension, excluding text extraction quality failures. These articles are not included in the primary training corpus but are retained in a separate supplementary partition that may be drawn upon for specific purposes, including the evaluation of model robustness to lower-quality training data, the investigation of domain coverage gaps identified in the primary corpus, and the generation of negative examples for tasks where the ability to distinguish high-quality from low-quality literature is itself an objective.

Articles classified as Excluded are those that achieve a low tier assignment on the relevance dimension, a low tier assignment on text extraction quality, or low tier assignments on both quality and suitability simultaneously. Excluded articles are removed from all training and supplementary partitions and their exclusion is recorded in the exclusion log with the specific combination of dimensional tier assignments and indicator scores that triggered the exclusion decision. The exclusion log is maintained as a component of the traceability system described in Section 6 and provides the evidential basis for the exclusion statistics reported in Section 4.

Ranking within tiers

Within the Core and Flagged corpus classifications, articles are further ranked to support prioritisation decisions in contexts where the available computational resources for a training run are insufficient to include the full corpus. The ranking is computed as a weighted sum of the nine individual indicator scores, with each indicator score converted to a numerical value of two for high, one for moderate, and zero for low, and each indicator assigned a weight reflecting its relative importance to training quality as determined during the calibration phase. The indicator weights are documented in the assessment scorecard provided in Annex A.

The resulting ranking score for each article ranges from zero to a maximum of eighteen, corresponding to the sum of the maximum weighted indicator scores across all three dimensions. Articles with identical ranking scores are further ordered by their relevance dimension tier, on the grounds that relevance to the project's target domains is the most direct determinant of the article's contribution to the specific extraction capabilities that the AIDESL platform is designed to develop. The ranking score is recorded as a metadata field in each article's assessment record and is available to the training pipeline as an input to any curriculum learning or data selection strategy that the WP2 team chooses to apply.

Handling of borderline cases

The decision rules described above are designed to produce unambiguous classifications for the large majority of articles in the corpus. A residual proportion of articles, however, present dimensional profiles that fall at the boundary between two classifications in ways that the automated scoring procedure cannot resolve with sufficient confidence. These borderline cases arise primarily in two situations: articles for which one or more indicator scores were assigned at the boundary between the moderate and low tiers by the automated procedure, and articles for which the automated and human review scores disagree on one or more indicators.

Borderline cases were resolved through a structured adjudication procedure in which two members of the Bosonit S.L. technical team independently reviewed the article's full dimensional profile and the specific indicator scores responsible for the borderline classification, and reached a consensus decision on the appropriate overall corpus classification. Where consensus could not be reached, the more

conservative classification was applied, consistent with the general principle that the integrity of the training corpus takes precedence over maximising its size. All adjudication decisions are recorded in the assessment record with a flag identifying them as human-adjudicated, enabling their retrospective analysis as part of the corpus quality evaluation reported in D7.6.

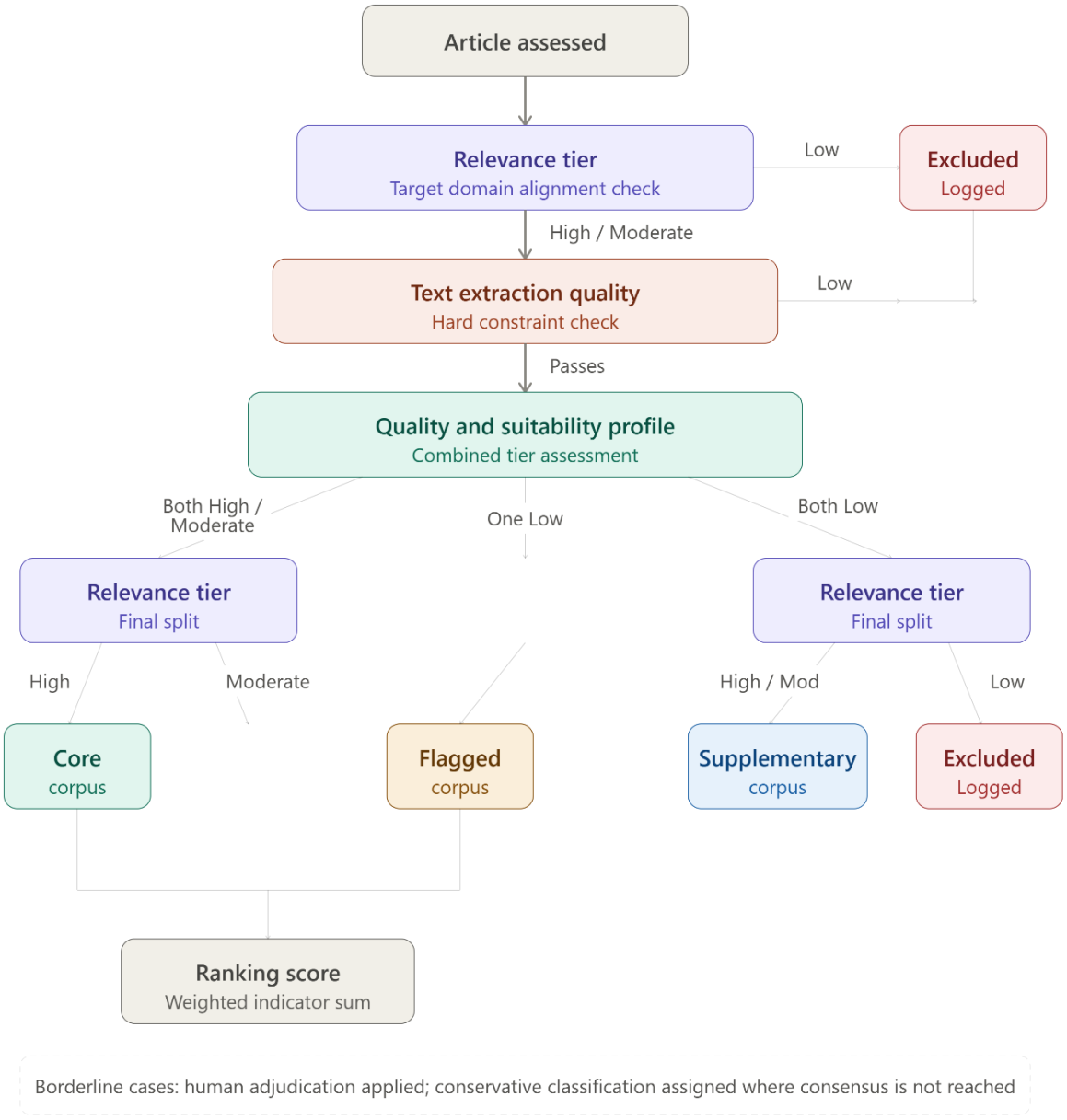

Figure 2. Decision flow for overall corpus classification based on dimensional tier assignments.

This figure reflects the unweighted maximum; the weighted maximum applied for internal ranking purposes is 32, as documented in Annex A."

4 Assessment Findings

This section presents the findings of the assessment process applied to the AIDESL corpus following the methodology described in Section 3. The findings are organised to move from the general to the specific: Section 4.1 provides an overall characterisation of the corpus in terms of its volume, composition, and distribution across the assessment tiers defined in Section 3.5; Section 4.2 disaggregates the findings by repository of origin; Section 4.3 presents the findings by target application domain and document type; and Section 4.4 identifies the principal gaps and limitations observed in the assembled literature. Together, these findings constitute the empirical foundation upon which the categorisation strategy defined in Section 5 was designed and the recommendations set out in Section 7 were formulated.

4.1 Overall corpus characterisation

The assessment process was applied to a total of 412 articles retrieved from the three selected repositories during the collection phase described in Section 2. This figure represents the number of articles that entered the assessment pipeline following the application of the initial eligibility filters established in D7.1, which excluded articles on the basis of licensing incompatibility, document type ineligibility, or temporal scope. The 412 articles therefore constitute the assessed corpus, from which the training corpus partitions described below were derived through the application of the scoring and ranking approach documented in Section 3.5.

The overall distribution of the assessed corpus across the four classification tiers is as follows. A total of 198 articles were assigned to the Core corpus classification, representing 48.1 percent of the assessed corpus. These articles achieved high or moderate tier assignments across all three assessment dimensions, with at least two high tier assignments, and constitute the primary training dataset for the LLM development activities of Work Package 2. A further 89 articles were assigned to the Flagged corpus classification, representing 21.6 percent of the assessed corpus. These articles were included in the training corpus with one or more pipeline-specific processing flags activating targeted preprocessing interventions, as detailed in D7.4. The combined Core and Flagged corpus therefore encompasses 287 articles, representing 69.7 percent of the assessed corpus, which constitutes the active training dataset available to Work Package 2 at the time of this deliverable's submission.

A total of 54 articles were assigned to the Supplementary corpus classification, representing 13.1 percent of the assessed corpus. These articles were retained in a separate partition for the purposes described in Section 3.5 but are not included in the primary training dataset. The remaining 71 articles, representing 17.2 percent of the assessed corpus, were assigned to the Excluded classification and removed from all training and supplementary partitions. Their exclusion decisions are documented in the exclusion log maintained as part of the traceability system described in Section 6, with the specific combination of dimensional tier assignments responsible for each exclusion recorded against the article's DOI.

Among the 71 excluded articles, the most frequent exclusion trigger was a low tier assignment on the relevance dimension, which accounted for 38 exclusions representing 53.5 percent of the total exclusions. These articles were retrieved by the ingestion pipeline as a result of broad query terms that captured content adjacent to the project's target domains but not sufficiently aligned with any of the three application domains to justify inclusion. A further 21 exclusions, representing 29.6 percent, were triggered by a low score on the text extraction quality indicator, which acted as a hard constraint as described in Section 3.5. These articles were predominantly PDFs whose text layer was either substantially incomplete due to rendering issues inherent in their source files, or so heavily

contaminated by non-prose content that the extractable text fell below the minimum threshold defined in the assessment scorecard. The remaining 12 exclusions, representing 16.9 percent, were triggered by low tier assignments on both the scientific quality and suitability dimensions simultaneously, in cases where the relevance tier assignment was insufficient to compensate for the combined deficiencies on the remaining dimensions.

The assessment process also identified a significant operational challenge in the automated retrieval phase that had material consequences for the composition of the assessed corpus. The systematic and high-volume nature of the programmatic article retrieval operations conducted during the initial bulk collection phase resulted in the temporary suspension of API access by two of the three selected repositories, which implemented rate-limiting and access restriction mechanisms in response to the volume and frequency of retrieval requests generated by the pipeline. This situation required the adoption of alternative retrieval strategies for the affected repositories, including the introduction of request throttling parameters, the use of staged retrieval operations distributed across extended time windows, and in a subset of cases the manual download and subsequent integration of articles that could not be retrieved programmatically within the operational timeframe. The access restriction episodes introduced delays into the collection timeline and resulted in gaps in the temporal and thematic coverage of the corpus that would otherwise have been filled by articles available in the restricted repositories during the affected periods. These gaps are identified and discussed further in Section 4.4, and their implications for the completeness of the training corpus are addressed in the recommendations set out in Section 7.

A governance-level finding of note is that approximately 30 percent of the articles processed through the assessment pipeline, corresponding to approximately 124 articles from the assessed corpus, triggered the sensitive data detection mechanism described in Section 6 and were routed through the human review and quarantine workflow prior to their corpus classification being finalised. The primary trigger for sensitive data flags across the assessed corpus was the presence of email addresses in the article text, which are classified as personal data under the RGPD framework established in D7.1 and therefore require human review before the article containing them can be admitted to the corpus. Email addresses appear with significant frequency in biomedical preprints, primarily in the context of corresponding author contact details embedded in the article body rather than confined to the journal metadata, and their systematic detection and handling represented a substantial component of the human review workload during the assessment period. Secondary triggers included the presence of identifiable patient identifiers in case report appendices and the detection of institutional affiliation strings that coincided with patterns associated with sensitive organisational data. The outcomes of the human review process for flagged articles are documented in the traceability logs described in Section 6, and the distribution of flag types across the assessed corpus is summarised in Annex C.

The ranking scores computed for Core and Flagged corpus articles, as described in Section 3.5, produced a distribution that reflects the heterogeneous quality profile of the preprint literature sources used for the project. Among Core corpus articles, the median ranking score was 14 out of a maximum of 18, with a distribution concentrated in the range of 12 to 16. Among Flagged corpus articles, the median ranking score was 9, with a broader distribution reflecting the greater variability in dimensional profiles that characterises this classification tier. These ranking scores are recorded in the metadata of each article's assessment record and are available to the Work Package 2 team as inputs to any data selection or curriculum learning strategy applied during model training.

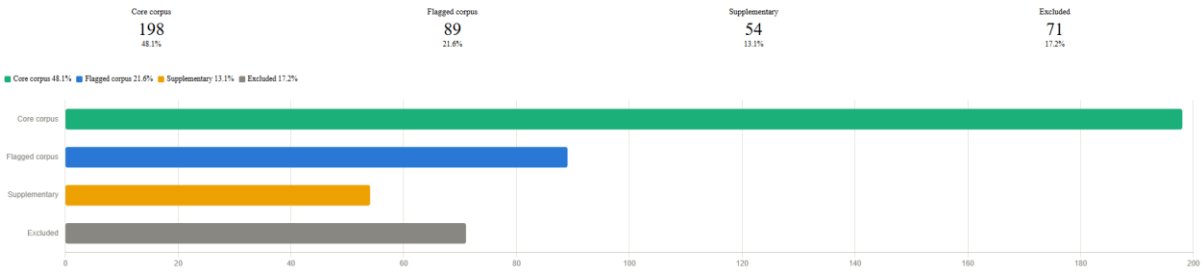

Figure 3. Distribution of the assessed corpus (n = 412) across the four classification tiers defined in the assessment framework.

4.2 Findings by repository

The 412 articles comprising the assessed corpus were drawn from the three selected repositories in proportions that reflect both the thematic scope of each source and the practical constraints encountered during the retrieval phase described in Section 2.2. arXiv contributed the largest share of the assessed corpus, accounting for approximately 42 percent of the total, equivalent to 173 articles. medRxiv contributed approximately 33 percent, equivalent to 136 articles, and bioRxiv contributed the remaining 25 percent, equivalent to 103 articles. The distribution across repositories is broadly consistent with the retrieval strategy described in Section 2.2, which prioritised arXiv as the primary source of methodological and AI-focused literature relevant to the technical development activities of the project while treating medRxiv and bioRxiv as the principal sources of domain-specific clinical and biological content.

The corpus classification outcomes differed meaningfully across the three repositories, reflecting the distinct thematic and structural characteristics of the literature deposited in each source.

Among the 173 articles sourced from arXiv, the proportion assigned to the Core corpus classification was the highest of the three repositories, with approximately 52 percent of arXiv articles achieving Core status. This outcome is attributable to the structural characteristics that distinguish arXiv preprints from those deposited in the two biomedical repositories. Articles in arXiv, particularly those in the quantitative biology and computer science subject areas most relevant to the AIDESL project, tend to exhibit a high degree of structural consistency and prose density that yields strong scores on the suitability for LLM training dimension. The prevalence of clearly delineated abstract, introduction, methodology, results, and discussion sections in arXiv computer science and quantitative biology submissions, combined with the typically high text extraction quality of PDFs generated from LaTeX source files, which are the dominant typesetting format in these communities, contributed positively to the suitability dimension scores of this repository's articles across the corpus. The exclusion rate for arXiv articles was correspondingly lower than for the other two repositories, at approximately 13 percent, with text extraction quality failures accounting for a smaller proportion of exclusions than in medRxiv and bioRxiv, where PDF rendering issues originating from journal submission formatting conventions were more prevalent.

Among the 136 articles sourced from medRxiv, the distribution across classification tiers was more heterogeneous. The Core corpus proportion for medRxiv articles was approximately 46 percent, slightly below the overall corpus average of 48.1 percent. The relevance dimension scores for medRxiv articles were generally strong, consistent with the repository's direct orientation towards clinical and health sciences content that maps well onto the three AIDESL target domains. The primary source of variability in medRxiv classification outcomes was the scientific quality dimension, where the absence of peer

review at the point of deposition was most consequential: a non-trivial proportion of medRxiv articles in the assessed corpus exhibited methodological reporting deficiencies on the reporting completeness indicator that resulted in moderate or low quality tier assignments. The exclusion rate for medRxiv articles was approximately 18 percent, with low relevance tier assignments accounting for a smaller proportion of exclusions than in arXiv, and text extraction quality failures and low scientific quality profiles accounting for a larger share.

Among the 103 articles sourced from bioRxiv, the Core corpus proportion was approximately 44 percent, the lowest of the three repositories. This outcome reflects the positioning of bioRxiv as a source of adjacent rather than central content for the AIDESL project: the repository's primary coverage of life sciences and biological research means that a larger proportion of its articles required assessment on the boundary between the high and moderate relevance tiers, with use case fit and topic specificity scores frequently insufficient to achieve a high composite relevance profile. The Flagged corpus proportion for bioRxiv articles was correspondingly higher than for the other two repositories, at approximately 25 percent, as a significant number of articles that cleared the relevance threshold did so with moderate rather than high tier assignments that triggered processing flags in the downstream pipeline. The exclusion rate for bioRxiv articles was approximately 19 percent, the highest of the three repositories, with low relevance tier assignments accounting for the majority of exclusions.

A finding that was consistent across all three repositories was the impact of the API access restriction episodes described in Section 4.1 on the coverage and completeness of the repository-level corpus partitions. Each of the three repositories implemented access restriction mechanisms in response to the volume of programmatic retrieval requests generated by the ingestion pipeline during the bulk collection phase, and the resolution strategies adopted for each were necessarily adapted to the specific technical and administrative constraints of the affected repository. For arXiv, the primary resolution strategy was the introduction of request throttling parameters within the ingestion pipeline, which reduced retrieval throughput but enabled continuous, if slower, access to the repository's content. For medRxiv and bioRxiv, which share an underlying infrastructure and implemented equivalent access restrictions simultaneously, the resolution required a combination of staged retrieval operations conducted across extended time windows and, in a subset of cases, the manual retrieval and subsequent integration of articles that could not be accessed programmatically within the operational constraints of the collection phase. The articles retrieved through manual intervention were processed through the same assessment pipeline as programmatically retrieved articles and are not distinguished in the corpus metadata, but their existence is noted here as a relevant contextual factor in interpreting the completeness of the repository-level findings.

The repository-level findings described in this section are summarised in the corpus statistics tables provided in Annex C, which present the full breakdown of classification tier assignments by repository alongside the associated exclusion trigger distributions.

4.3 Findings by domain and medical specialty

The thematic distribution of the assessed corpus across the three AIDESL target application domains reflects the retrieval strategy described in Section 2.2 and the relevance assessment outcomes documented in Section 4.1. Clinical research constitutes the dominant domain within the corpus, accounting for approximately half of the articles assigned to the Core and Flagged corpus classifications. Medical device development and pharmaceutical evaluation are represented in broadly comparable proportions, each accounting for approximately a quarter of the active training dataset. This distribution is consistent with the retrieval query design, which weighted clinical research terminology more heavily in recognition of its centrality to the project's primary demonstration scenarios, and with the differential

availability of open-access preprint literature across the three domains, which is more abundant in clinical research than in the more commercially sensitive areas of medical device development and pharmaceutical evaluation.

The clinical research domain encompasses the widest thematic range within the corpus, reflecting the breadth of the domain itself. Articles in this partition address a diverse set of research questions spanning epidemiological studies of disease incidence and prevalence, clinical trial reports evaluating therapeutic interventions across a range of conditions, outcomes research examining the real-world effectiveness of established treatments, and health services research addressing the organisation and delivery of clinical care. This thematic diversity is a deliberate feature of the corpus design rather than an incidental outcome of the retrieval process: the AIDESL platform is intended to support the extraction of structured data from SLRs across the full spectrum of clinical research, and a corpus that is too narrowly focused on a specific clinical subdomain would produce models with insufficient generalisability to support the project's target use cases. The breadth of the clinical research partition therefore reflects a conscious trade-off between depth of coverage within specific subdomains and breadth of coverage across the domain as a whole, with the balance tilted towards breadth at this stage of corpus development.

The pharmaceutical evaluation domain partition is characterised by a higher concentration of methodologically homogeneous articles relative to the clinical research partition. The majority of pharmaceutical evaluation articles in the corpus address pharmacokinetic and pharmacodynamic properties of therapeutic compounds, clinical trial outcomes for drug candidates at various stages of development, and comparative effectiveness analyses of established pharmaceutical treatments. The relative homogeneity of this partition reflects both the more constrained thematic scope of pharmaceutical research as represented in preprint repositories and the specificity of the retrieval queries applied to this domain, which were designed to target the types of pharmaceutical literature most directly relevant to the data extraction tasks of the AIDESL demonstration scenarios. The exclusion rate within the pharmaceutical evaluation domain was marginally higher than the corpus-wide average, driven primarily by a higher prevalence of articles reporting preliminary or early-phase findings with limited statistical context that resulted in low tier assignments on the evidence level indicator of the scientific quality dimension.

The medical device development domain presents the most distinctive profile within the corpus in terms of both thematic content and assessment outcomes. This domain encompasses articles addressing the technical development and validation of medical devices, clinical evaluations of device performance and safety, regulatory science contributions relevant to device approval pathways, and methodological studies of the assessment frameworks applied to medical devices. The medical device literature available through the three selected preprint repositories is sparser than that available for the other two domains, a reflection of the fact that medical device research is more heavily concentrated in peer-reviewed specialist journals and conference proceedings that fall outside the open-access preprint ecosystem targeted by the project's collection strategy. This scarcity of available literature had a direct consequence for the corpus: the medical device partition of the active training dataset is proportionally underrepresented relative to the domain's importance to the project's demonstration scenarios, a gap that is identified as a priority recommendation for the next phase of corpus development in Section 7.

Turning to document type, the assessed corpus is dominated by primary research articles, which account for the large majority of articles across all three domains and all three repositories. This outcome is a direct consequence of the document type eligibility criteria applied during the collection phase, which excluded editorial content, correspondence, opinion pieces, and commentary from the corpus at the point of ingestion, and of the document type distribution characteristic of preprint

repositories, in which primary research articles represent the overwhelming majority of deposited content.

Systematic reviews and meta-analyses are present in the corpus but constitute a secondary rather than primary document type, representing a meaningful minority of articles across the three domains. Their presence in the corpus warrants specific comment given the thematic positioning of the AIDESL project. The platform under development is designed precisely to automate the data extraction processes that underpin systematic literature reviews, and systematic reviews themselves therefore represent a document type of particular strategic value as training material: they provide structured, high-density representations of the biomedical evidence base that are rich in the types of structured information that the platform's extraction models are trained to identify and extract. Systematic reviews and meta-analyses in the corpus received uniformly strong scores on the document structure coherence indicator of the suitability dimension, consistent with the high degree of structural standardisation characteristic of this document type, and their Core corpus classification rate was above the corpus-wide average. Their relatively modest representation in the corpus is attributable to their lower prevalence in preprint repositories relative to primary research articles, rather than to any systematic disadvantage in the assessment process.

Clinical trial reports constitute a further document type of note within the corpus, particularly within the clinical research and pharmaceutical evaluation domain partitions. These articles share the structural regularity of systematic reviews but are more abundant in the preprint repositories and provide a complementary type of structured information that is directly relevant to the extraction tasks of the AIDESL demonstration scenarios. Their assessment outcomes were broadly similar to those of systematic reviews on the suitability dimension, with strong document structure coherence scores reflecting the CONSORT reporting standard that governs clinical trial reporting, and variable scientific quality scores reflecting the range of methodological rigour represented across trials at different stages of development and with different sample sizes and endpoint definitions.

4.4 Identified gaps and limitations

The assessment process, in addition to producing the corpus classification outcomes described in Sections 4.1 through 4.3, generated a set of findings concerning the gaps and limitations of the assembled corpus that have direct implications for the downstream activities of Work Package 7 and for the LLM development work of Work Package 2. These findings are documented in this section both to provide a transparent account of the current state of the corpus and to inform the recommendations set out in Section 7, which address the mitigation strategies proposed for each identified gap.

The most significant structural gap in the assessed corpus concerns the underrepresentation of medical device development literature relative to the domain's importance to the project's demonstration scenarios. As noted in Section 4.3, the open-access preprint ecosystem targeted by the project's collection strategy does not adequately reflect the volume and diversity of medical device research produced within the broader scientific community. Medical device research is disproportionately concentrated in peer-reviewed specialist journals, conference proceedings, and regulatory submission documents that fall outside the scope of the three selected repositories, and the retrieval strategy applied during the bulk collection phase was consequently unable to assemble a medical device partition of sufficient depth to support the full range of extraction tasks envisaged for this domain in the project's demonstration scenarios. The practical consequence of this gap is that the training corpus available to Work Package 2 for the medical device domain is less comprehensive than for clinical research and pharmaceutical evaluation, and the extraction models trained on the current corpus may exhibit lower performance on medical device-specific extraction tasks than on tasks associated with the

better-represented domains. This gap is not resolvable within the current source portfolio and collection strategy, and its mitigation will require either the identification of additional open-access sources specifically covering medical device research, or the negotiation of access arrangements for relevant subscription-based content that comply with the licensing framework established in D7.1.

A second gap concerns the temporal and thematic coverage discontinuities introduced by the API access restriction episodes documented in Sections 4.1 and 4.2. The periods during which programmatic access to the three repositories was suspended, and the staged retrieval operations adopted as resolution strategies, resulted in gaps in the corpus that are not randomly distributed across the thematic and temporal dimensions of the literature. The articles that could not be retrieved during the access restriction periods are precisely those that were published or deposited within the affected repositories during those periods, and their absence from the corpus means that certain recent thematic developments within the covered domains may be underrepresented relative to their actual prevalence in the literature. The manual retrieval operations conducted to partially compensate for these gaps were effective in recovering a subset of the affected articles but could not provide complete coverage given the operational constraints of the collection phase. The net effect is a corpus whose temporal coverage is uneven in ways that are not fully transparent from the corpus metadata alone, since the articles successfully retrieved through manual intervention are not distinguished from those retrieved programmatically. This limitation is acknowledged here as a relevant factor in interpreting the temporal distribution statistics reported in Annex C.

A third gap concerns the representation of systematic reviews and meta-analyses within the corpus relative to their strategic value as training material for the AIDESL platform. As discussed in Section 4.3, this document type is present in the corpus but constitutes a secondary rather than primary component, reflecting its lower prevalence in preprint repositories rather than any deficiency in the assessment process. The gap is consequential because systematic reviews and meta-analyses provide the most structurally regular and semantically dense representations of the biomedical evidence base available in the scientific literature, and their underrepresentation in the training corpus relative to primary research articles means that the extraction models developed in Work Package 2 will be trained on a corpus whose structural characteristics do not fully reflect the document type that the platform is ultimately designed to process. Addressing this gap will require a targeted supplementary collection effort focused specifically on systematic review and meta-analysis content, potentially drawing on sources beyond the three repositories selected for the primary collection phase, subject to the licensing constraints established in D7.1.

Beyond these three primary gaps, the assessment process identified a set of more diffuse limitations that, while individually less consequential than the structural gaps described above, collectively represent a relevant qualification on the comprehensiveness and representativeness of the corpus. These include the exclusive use of English-language articles, which reflects an operational constraint of the processing pipeline but introduces a systematic bias towards anglophone research traditions and may underrepresent contributions from research communities whose primary publication language is not English. They also include the concentration of the corpus in the period from 2020 onwards, which, while justified by the considerations described in Section 2.2, means that the corpus does not fully capture methodological developments and evidence accumulated prior to the accelerated growth of preprint deposition in the biomedical domain. Finally, the corpus is limited to articles available through the public APIs of the three selected repositories at the time of the ingestion operations, and does not include articles that were deposited after those operations or that were available in the repositories but not returned by the retrieval queries due to the specificity of the query terms applied.

These gaps and limitations are not presented as fundamental deficiencies in the corpus but as the documented constraints of a first-phase collection effort conducted within the timeline, resource, and licensing boundaries of Work Package 7. The corpus as assembled provides a substantive and well-characterised foundation for the LLM development activities of Work Package 2, and the gaps identified here represent a prioritised agenda for corpus expansion in subsequent phases of the project rather than an obstacle to the current phase of model development. The specific recommendations arising from these findings are set out in Section 7.

5 Categorization Strategy

The categorization strategy developed under Task 7.3 addresses the challenge of organising the assessed corpus in a manner that is simultaneously useful for LLM training, compatible with the semantic interoperability requirements of the federated health data space described in D7.5, and sufficiently transparent and reproducible to support the auditability obligations established in D7.1. Categorization in this context serves a different function from the assessment process described in Section 3: where assessment determines whether and how an article is admitted to the corpus, categorization determines how admitted articles are organised, labelled, and made accessible to the downstream processes that consume the corpus. The two processes are sequential and complementary, with the categorization strategy applied to the subset of articles that have cleared the assessment pipeline and been assigned to the Core, Flagged, or Supplementary corpus classifications.

The strategy rests on two complementary layers of organisation. The first layer is a domain taxonomy that assigns each article to one or more of the three AIDESL target application domains, providing a coarse-grained organisational structure that maps directly onto the project's demonstration scenarios and enables domain-specific filtering of the corpus for training and evaluation purposes. The second layer is a semantic labelling scheme based on Medical Subject Headings controlled vocabulary, which provides a fine-grained, standardised, and internationally recognised representation of the thematic content of each article at the level of individual concepts, enabling interoperability with external biomedical knowledge resources and supporting the semantic retrieval and filtering operations required by the federated data space architecture.

These two layers are designed to be complementary rather than redundant. The domain taxonomy provides the coarse-grained organisational structure that maps most directly onto the project's operational needs, while the MeSH labelling layer provides the granularity and standardisation required for interoperability and semantic precision. Together they constitute a categorization scheme that is both practically useful within the immediate context of the AIDESL project and extensible to the broader requirements of the health data space infrastructure described in D7.5.

5.1 Categorization objectives and design principles

The design of the categorization strategy was guided by four objectives that were defined by Bosonit S.L. at the outset of Task 7.3 and validated against the technical requirements of the downstream work packages throughout the development process.

The first objective was operational utility for LLM training. The categorization scheme needed to produce organisational labels that are directly actionable by the training pipeline developed under Work Package 2, enabling domain-specific training runs, targeted evaluation of model performance across application domains, and the application of curriculum learning strategies that exploit the thematic structure of the

corpus. This objective imposed a requirement for simplicity and consistency in the domain taxonomy layer: a categorization scheme with a large number of fine-grained categories would provide richer organisational structure but would introduce complexity that the training pipeline could not exploit without substantial additional development effort. The decision to anchor the domain taxonomy on the three AIDESL target application domains reflects this objective directly.

The second objective was semantic interoperability. The categorization scheme needed to produce labels that are compatible with the semantic standards and knowledge resources used in the broader biomedical information ecosystem, enabling the corpus metadata to be integrated with external resources and consumed by the federated data space infrastructure without requiring bespoke translation or mapping procedures. This objective motivated the adoption of MeSH as the primary controlled vocabulary for the semantic labelling layer, given its position as the most widely adopted standard for biomedical literature indexing at the international level and its comprehensive coverage of the thematic areas relevant to the three AIDESL target domains.

The third objective was reproducibility and consistency. The categorization process needed to produce results that are consistent across articles assessed at different points in time, by different members of the team, and through different combinations of automated and manual procedures. This objective imposed requirements on the design of the categorization tools and workflows: the automated MeSH term extraction procedure needed to produce stable outputs for a given article regardless of when it was run, and the domain taxonomy assignment rules needed to be specified with sufficient precision to minimise the scope for inconsistent application. Reproducibility was particularly important in the context of the project's auditability obligations, since the categorization labels assigned to each article form part of its permanent metadata record and must be traceable to a consistent and documented procedure.

The fourth objective was extensibility. The categorization scheme needed to be designed in a manner that would accommodate the expansion of the corpus in subsequent phases of the project without requiring fundamental revision of the organisational structure or the semantic labelling approach. This objective was addressed primarily through the choice of MeSH as the semantic labelling vocabulary, which is maintained by the National Library of Medicine and updated annually to reflect developments in biomedical knowledge and terminology, ensuring that the labelling scheme remains current and extensible as the corpus grows. The domain taxonomy, by contrast, is fixed to the three AIDESL target domains for the duration of the project, reflecting the stability of the project's application scope.

Four design principles followed from these objectives and were applied consistently throughout the development of the categorization strategy.

The first design principle was separation of layers. The domain taxonomy and the MeSH semantic labelling scheme were designed and implemented as independent layers of the categorization metadata, each serving its own purpose and neither dependent on the other for its integrity. An article's domain taxonomy assignment does not determine or constrain its MeSH labels, and vice versa. This separation ensures that each layer can be updated, extended, or replaced independently without affecting the other, and that downstream processes can consume either layer in isolation or both in combination according to their specific requirements.

The second design principle was tolerance of multi-domain assignment. Many articles in the biomedical literature address topics that span more than one of the three AIDESL target domains: a study evaluating the clinical performance of a pharmaceutical treatment delivered through a medical device, for instance, is simultaneously relevant to all three domains. The domain taxonomy was designed to accommodate this reality by allowing articles to be assigned to multiple domains where justified by their content, rather than forcing a single-domain assignment that would misrepresent the thematic scope of multi-domain

articles and reduce the utility of the corpus for cross-domain training tasks. The proportion of articles with multi-domain assignments in the assessed corpus is reported in Annex C.

The third design principle was primacy of automated procedures with human validation. The scale of the assessed corpus made fully manual categorization impractical within the resources available to Task 7.3, and the categorization strategy was therefore designed from the outset to rely primarily on automated procedures, specifically the programmatic MeSH term extraction described in Section 5.3 and the domain classification heuristics described in Section 5.2, with human validation reserved for a stratified sample of the corpus and for cases where the automated procedures returned ambiguous or low-confidence results. This principle reflects a deliberate trade-off between the accuracy that full manual categorization would provide and the scalability that the project requires as the corpus expands in subsequent phases.

The fourth design principle was metadata persistence. All categorization outputs, including domain taxonomy assignments, MeSH term lists, confidence scores associated with automated assignments, and human validation flags, are stored as persistent metadata fields in each article's Silver layer object within the pipeline infrastructure. These metadata fields are immutable once finalised, ensuring that the categorization record for each article remains stable and auditable regardless of subsequent changes to the corpus or the categorization procedure. Updates to categorization assignments resulting from revised automated procedures or human review decisions are recorded as new metadata versions rather than overwriting the original assignment, preserving the full history of each article's categorization record.

5.2 Taxonomy definition

The domain taxonomy constitutes the first and coarser of the two categorization layers applied to the assessed corpus. Its function is to assign each article in the Core, Flagged, and Supplementary corpus classifications to one or more of the three AIDESL target application domains, producing a domain label that serves as the primary organisational axis of the corpus and the principal filter applied by downstream training and evaluation processes.

The three domains of the taxonomy are defined as follows. The clinical research domain encompasses articles addressing the design, conduct, reporting, and synthesis of research into the safety, efficacy, and effectiveness of health interventions in human populations, including clinical trials, observational studies, outcomes research, epidemiological investigations, and systematic reviews of clinical evidence. The pharmaceutical evaluation domain encompasses articles addressing the development, characterisation, and clinical assessment of pharmaceutical compounds and biological therapies, including pharmacokinetic and pharmacodynamic studies, drug safety and tolerability evaluations, comparative effectiveness analyses of pharmaceutical treatments, and regulatory science contributions relevant to drug approval. The medical device development domain encompasses articles addressing the technical development, performance validation, clinical evaluation, and regulatory assessment of medical devices, diagnostic tools, and health technologies, including studies of device safety and efficacy, human factors research relevant to device design, and methodological contributions to the assessment frameworks applied to medical devices.

These three domain definitions are intentionally broad, reflecting the need to accommodate the thematic diversity of biomedical preprint literature within a taxonomy of manageable complexity. The breadth of each domain definition means that the domain taxonomy functions as a coarse-grained organisational tool rather than a precise thematic classification, and that the granularity required for more specific thematic filtering is provided by the MeSH semantic labelling layer described in Section 5.3 rather than by

the domain taxonomy itself. This division of labour between the two categorization layers is consistent with the design principles set out in Section 5.1.

Domain taxonomy assignments were produced through an automated classification procedure applied to the title, abstract, and section headings of each article's extracted text. The procedure applied the same keyword-based matching mechanism described in Section 3.2 in the context of the relevance assessment, using a controlled vocabulary of domain-specific terms derived from MeSH thematic branches and refined iteratively during the calibration phase of the assessment process. For each of the three domains, the classification procedure computed a domain relevance signal based on the density and distribution of domain-specific terms across the article's title, abstract, and headings, and assigned a domain label where the signal exceeded a defined threshold. Articles for which the domain relevance signal exceeded the threshold for more than one domain were assigned multiple domain labels, consistent with the multi-domain assignment principle described in Section 5.1.

The threshold values applied to the domain relevance signals were calibrated during the same calibration exercise used to validate the assessment scorecard, using a stratified sample of articles whose domain assignments had been independently verified by two members of the Bosonit S.L. technical team. The calibration exercise produced threshold values that achieved satisfactory agreement with the human-verified domain assignments across the calibration sample, with the highest agreement rates observed for articles in the clinical research domain, where the terminological specificity of the relevant MeSH branches provided a strong and consistent classification signal, and the lowest agreement rates observed for articles at the boundary between the pharmaceutical evaluation and clinical research domains, where the overlap between domain-specific terminologies introduced ambiguity that the keyword-based procedure could not fully resolve. Articles for which the automated classification returned a signal below the threshold for all three domains, or for which the signal fell within a defined ambiguity band around the threshold for one or more domains, were escalated to human review for manual domain assignment.

The distribution of the assessed corpus across the three domain taxonomy categories, including the proportion of articles with multi-domain assignments and the breakdown of domain assignments by corpus classification tier, is presented in Annex C.

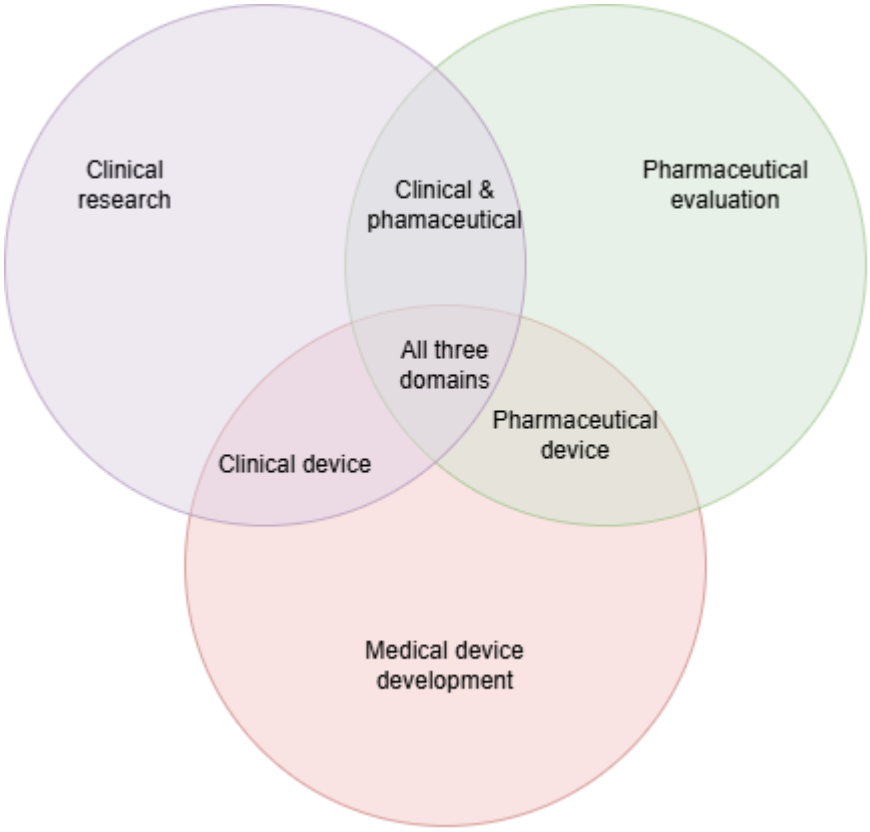

Figure 4. Conceptual representation of the AIDESL domain taxonomy showing the three target application domains and their overlap zones for articles with multi-domain assignments.

5.3 Semantic labelling with MeSH

The Medical Subject Headings controlled vocabulary, maintained by the National Library of Medicine of the United States and updated on an annual basis, constitutes the semantic labelling standard applied to all articles in the Core, Flagged, and Supplementary corpus classifications. MeSH is the most widely adopted controlled vocabulary for the indexing of biomedical literature at the international level, providing a hierarchically organised set of descriptors that cover the full thematic range of biomedical and health sciences research. Its adoption as the primary semantic labelling standard for the AIDESL corpus reflects both its thematic comprehensiveness and its position as the de facto interoperability standard for biomedical literature metadata, which ensures that the semantic labels assigned to corpus articles are compatible with external knowledge resources and retrieval systems without requiring bespoke mapping or translation procedures.

MeSH term extraction was implemented as an automated procedure integrated into the corpus processing pipeline developed under Task 7.3. The procedure consumes the extracted text of each article, specifically its title and abstract, and submits this content to the MeSH term extraction API provided by the National Library of Medicine as part of its publicly accessible suite of biomedical text mining services. The API applies the NLM Medical Text Indexer, a machine learning system trained on the full corpus of MEDLINE-indexed literature, to identify and return the MeSH descriptors most relevant to

the submitted text. For each article processed through the procedure, the API returns a ranked list of MeSH descriptors accompanied by confidence scores that reflect the system's estimated probability that each descriptor accurately characterises the article's thematic content. The full ranked list of descriptors and their associated confidence scores are stored as persistent metadata fields in each article's Silver layer object, providing a complete and auditable record of the semantic labelling output for every article in the corpus.

The MeSH vocabulary is organised into a hierarchical tree structure comprising sixteen broad thematic branches, each of which is subdivided into progressively more specific descriptor levels. The descriptors returned by the NLM API for a given article may span multiple branches of the MeSH tree, reflecting the multidisciplinary character of biomedical research and the tendency of individual articles to address topics that draw on concepts from more than one thematic area. For the purposes of the AIDESL corpus, no restriction was imposed on the branches from which descriptors could be drawn, and the full range of MeSH descriptors returned by the API for each article was retained in the article's metadata. This approach ensures that the semantic labelling captures the full thematic profile of each article rather than a subset filtered by branch relevance, and that the metadata is available to support a wider range of retrieval and filtering operations than would be possible with a branch-restricted labelling scheme.

In operational terms, the MeSH labelling procedure was applied to each article at the point of its admission to the corpus following the completion of the assessment process. Articles assigned to the Core and Flagged corpus classifications were labelled as part of the standard pipeline processing sequence, with the MeSH extraction step executed automatically following the Silver layer transformation described in the WP7 pipeline documentation. Articles assigned to the Supplementary corpus classification were also labelled through the same procedure, ensuring that the supplementary partition is semantically characterised to the same standard as the active training dataset and can be integrated into the primary corpus in future phases without requiring retrospective labelling. The MeSH labelling procedure was not applied to excluded articles, which were removed from the pipeline prior to the labelling step and whose metadata records contain only the assessment scores and exclusion trigger information documented in the exclusion log.

The volume and distribution of MeSH descriptors across the assessed corpus reflects the thematic breadth and diversity of the literature assembled for the project. Articles in the clinical research domain partition tend to attract a larger number of MeSH descriptors per article than those in the pharmaceutical evaluation or medical device development partitions, consistent with the greater thematic complexity and multidisciplinary character of clinical research literature. The MeSH descriptors most frequently assigned across the full corpus are concentrated in the Diseases, Analytical, Diagnostic and Therapeutic Techniques and Equipment, and Health Care branches of the MeSH tree, reflecting the clinical orientation of the corpus as a whole. A full listing of the MeSH descriptors applied across the corpus, organised by frequency of assignment and by domain partition, is provided in Annex B.

The integration of the MeSH labelling layer with the domain taxonomy described in Section 5.2 enables a two-dimensional navigation of the corpus that is more powerful than either layer alone. A user of the corpus metadata can filter articles by domain taxonomy assignment to retrieve all articles relevant to a given AIDESL application domain, and can then apply MeSH descriptor filters within that domain partition to identify articles addressing specific thematic subsets of the domain. Conversely, a user can begin with a MeSH descriptor of interest and retrieve all articles in the corpus to which that descriptor has been assigned, irrespective of their domain taxonomy classification, enabling cross-domain thematic retrieval that is not constrained by the coarser-grained domain boundaries. This two-dimensional navigation capability is of particular value for the data selection and curriculum learning operations that the Work

Package 2 team may wish to apply during model training, and for the semantic retrieval operations that the federated data space architecture described in D7.5 will need to support.

5.4 Categorization process and tools

The categorization process applied to the AIDESL corpus was implemented as a sequential two-stage procedure, with the domain taxonomy assignment and the MeSH semantic labelling executed in that order for each article admitted to the corpus following the completion of the assessment process described in Section 3. The two stages share a common technical foundation in the text extraction and Silver layer transformation pipeline developed under Task 7.3, but draw on distinct tools and produce distinct metadata outputs that are stored independently in each article's metadata record.

The first stage of the categorization process is domain taxonomy assignment. As described in Section 5.2, this stage applies a keyword-based classification procedure to the title, abstract, and section headings of each article's extracted text, computing a domain relevance signal for each of the three AIDESL target domains and assigning domain labels where the signal exceeds the calibrated threshold values. This procedure is implemented as a processing module within the WP7 pipeline, executed automatically as part of the Silver layer transformation sequence that follows the initial text extraction step. The module consumes the structured text output of the extraction procedure and produces a domain assignment record containing the domain labels assigned to the article, the relevance signal values computed for each domain, and a flag indicating whether the assignment was produced by the automated procedure or by human review. This record is written to the article's Silver layer metadata object and persists unchanged through all subsequent pipeline stages.

Articles for which the automated domain assignment procedure returns a signal below the defined threshold for all three domains, or a signal within the ambiguity band around the threshold for one or more domains, are flagged for human review. The human review procedure for domain assignment involves the manual inspection of the article's title, abstract, and, where necessary, its methods and results sections by a member of the Bosonit S.L. technical team, who applies the domain definitions set out in Section 5.2 to determine the appropriate domain label or labels. The outcome of the human review is recorded in the domain assignment record alongside a flag identifying the assignment as human-reviewed, and the article proceeds to the second stage of the categorization process without further automated intervention on the domain assignment.

The second stage of the categorization process is MeSH semantic labelling. As described in Section 5.3, this stage submits the title and abstract of each article to the NLM Medical Text Indexer API and stores the ranked list of MeSH descriptors returned by the API, together with their associated confidence scores, in the article's Silver layer metadata object. The API call is implemented as a separate processing module within the pipeline, executed sequentially after the domain taxonomy assignment module and consuming the same structured text output. The module applies a configurable confidence score threshold to filter the descriptors returned by the API, retaining only those descriptors whose confidence score meets or exceeds the threshold and discarding lower-confidence descriptors that are unlikely to accurately characterise the article's thematic content. The threshold value applied during the current phase of corpus development was determined empirically during the calibration exercise described in Section 3.1, based on the agreement rate between API-returned descriptors above the threshold and the MeSH terms independently assigned by human reviewers to the calibration sample.

The MeSH labelling module produces a descriptor record for each article containing the filtered list of MeSH descriptors, their confidence scores, the API response timestamp, and the version of the NLM Medical Text Indexer applied at the time of the call. The version identifier is stored as a metadata field to

support the retrospective identification of any labelling inconsistencies that might arise from changes in the NLM system between API calls conducted at different points in the corpus development timeline, and to enable the targeted re-labelling of affected articles in the event that a significant change in API behaviour is detected.

The complete categorization metadata produced by the two-stage process, encompassing domain taxonomy assignments and MeSH descriptor records, is available to all downstream pipeline stages and to the external consumers of the corpus metadata, including the Work Package 2 training pipeline and the federated data space integration layer described in D7.5. The metadata is stored in a structured JSON format consistent with the Silver layer schema documented in the WP7 pipeline specifications and is accessible through the same query interface used to retrieve all other corpus metadata fields. A representative excerpt of the categorization metadata schema is provided in Annex B alongside the full MeSH descriptor listing.

The tooling underlying the categorization process was developed entirely within the WP7 pipeline infrastructure, with no dependency on external categorization platforms or third-party annotation tools beyond the NLM Medical Text Indexer API. This design choice reflects the principle of pipeline self-containment established during the technical architecture phase of Task 7.3, which prioritised the minimisation of external dependencies in order to reduce the operational risk associated with changes in the availability or behaviour of third-party services. The NLM API is the sole external dependency in the categorization process, and its selection was motivated by the considerations documented in Section 5.3, including its free accessibility, its stability as a service maintained by a public institution, and its direct suitability for the biomedical literature indexing task.

5.5 Corpus distribution by category

The application of the two-stage categorization process described in Section 5.4 to the 287 articles comprising the active training dataset, encompassing the Core and Flagged corpus classifications, produced a domain taxonomy distribution that reflects both the thematic orientation of the retrieval strategy described in Section 2.2 and the relevance assessment outcomes documented in Section 4. The Supplementary corpus partition of 54 articles was categorized through the same procedure and its distribution is reported separately below for completeness, though it does not form part of the active training dataset available to Work Package 2 at the time of this deliverable's submission.

Within the active training dataset of 287 articles, clinical research is the dominant domain, accounting for 148 articles representing 51.6 percent of the active training dataset. This proportion is consistent with the retrieval strategy, which weighted clinical research terminology more heavily in recognition of its centrality to the project's primary demonstration scenarios, and with the differential availability of open-access preprint literature across the three domains, as discussed in Section 4.3. Pharmaceutical evaluation accounts for 78 articles representing 27.2 percent of the active training dataset, and medical device development accounts for 47 articles representing 16.4 percent. The remaining 14 articles, representing 4.9 percent of the active training dataset, were assigned to more than one domain through the multi-domain assignment procedure described in Section 5.2, with the clinical research and pharmaceutical evaluation combination being the most frequent multi-domain profile, followed by clinical research and medical device development. No articles in the active training dataset were assigned to all three domains simultaneously, reflecting the thematic specificity required to meet the domain relevance signal threshold for all three domains concurrently.

Within the Core corpus classification specifically, comprising 198 articles, the domain distribution follows a similar pattern to the active training dataset as a whole, with clinical research accounting for

approximately 52 percent of Core articles, pharmaceutical evaluation for approximately 27 percent, and medical device development for approximately 15 percent, with the remaining articles carrying multi-domain assignments. Within the Flagged corpus classification of 89 articles, the medical device development domain accounts for a slightly higher proportion than in the Core corpus, consistent with the finding reported in Section 4.3 that medical device articles were more frequently assigned to the Flagged classification due to the combination of moderate relevance tier assignments and pipeline processing flags associated with the structural characteristics of device-focused literature.

Within the Supplementary corpus partition of 54 articles, the domain distribution differs more markedly from the active training dataset. Clinical research remains the dominant domain but at a lower proportion of approximately 44 percent, while articles with low relevance tier assignments that nonetheless cleared the exclusion threshold account for a higher share of the partition than in the active training dataset. The medical device development domain is proportionally better represented in the Supplementary partition than in the active training dataset, reflecting the higher exclusion rate and lower Core corpus classification rate of device-focused literature documented in Section 4.3, which resulted in a larger proportion of device articles being retained in the supplementary rather than active partition.

Turning to the MeSH semantic labelling distribution, the descriptor landscape across the active training dataset reflects the clinical and methodological orientation of the corpus described in preceding sections. The most frequently assigned MeSH descriptors are concentrated in the Diseases branch of the MeSH tree, particularly in the subdomains of cardiovascular diseases, neoplasms, and nervous system diseases, which are among the most heavily researched areas in the clinical literature available through the selected repositories. The Analytical, Diagnostic and Therapeutic Techniques and Equipment branch is the second most represented, with descriptors relating to clinical trial methodology, systematic review procedures, and diagnostic imaging techniques appearing with particular frequency across the clinical research and pharmaceutical evaluation domain partitions. The Health Care branch contributes a further significant share of the descriptor landscape, with descriptors addressing healthcare delivery, patient outcomes, and health services research appearing predominantly in the clinical research partition.

Within the pharmaceutical evaluation domain partition, the most frequently assigned MeSH descriptors are drawn from the Chemicals and Drugs branch of the MeSH tree, reflecting the compound-specific and pharmacological orientation of the literature in this partition. Within the medical device development partition, descriptors from the Equipment and Supplies subgroup of the Analytical, Diagnostic and Therapeutic Techniques and Equipment branch predominate, consistent with the technical and evaluative focus of device-related literature. The full listing of MeSH descriptors assigned across the corpus, organised by frequency of assignment and disaggregated by domain partition, is provided in Annex B.

The corpus distribution findings described in this section confirm that the categorization strategy has produced a well-characterised and navigable corpus that supports the downstream requirements of Work Package 2 and the federated data space integration activities of Work Package 7. The domain imbalance between clinical research and medical device development, identified as a gap in Section 4.4, is reflected directly in the distribution figures reported here and reinforces the recommendation set out in Section 7 for a targeted supplementary collection effort focused on medical device literature in the next phase of corpus development.

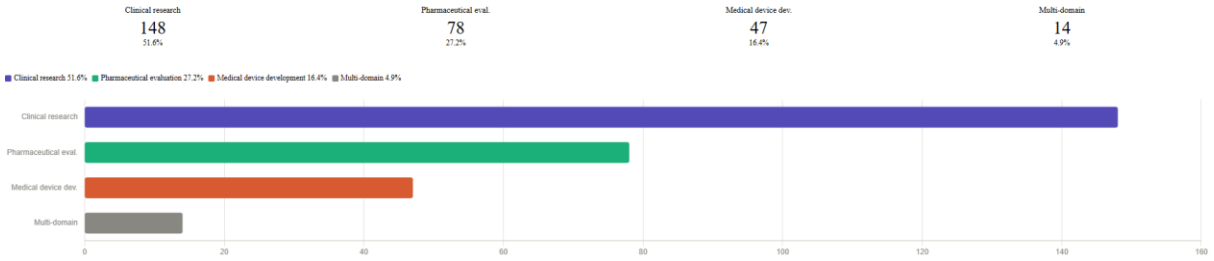

Figure 5. Distribution of the active training dataset (n = 287) across the AIDESL domain taxonomy categories.

6 Data Governance in the Assessment Process

The data governance mechanisms applied during the assessment process constitute an operational implementation of the compliance framework established in D7.1, adapted to the specific context of automated literature ingestion and corpus construction. While D7.1 defines the normative principles and regulatory obligations that govern all data collection and processing activities under Work Package 7, the present section documents how those principles were operationalised within the assessment pipeline: the technical systems deployed to detect potentially sensitive information in the ingested literature, the workflows through which flagged articles were routed for human review, and the traceability mechanisms that ensure every governance decision is recorded and auditable.

The governance layer described in this section operates as an integral component of the assessment pipeline rather than as a separate post-processing step. Its position within the pipeline sequence means that no article can progress from the raw ingestion layer to the corpus classification stage without having passed through the sensitive data detection procedure, ensuring that the governance obligations established in D7.1 are applied universally and consistently across the full assessed corpus. This design reflects a deliberate architectural choice to embed compliance controls within the pipeline infrastructure rather than applying them retrospectively, which would introduce both operational risk and auditability gaps that are incompatible with the transparency requirements of the EU AI Act and the data protection obligations of the RGPD.

6.1 Sensitive data detection during assessment

The sensitive data detection system implemented within the WP7 pipeline is responsible for identifying articles whose content contains personal data or other categories of sensitive information that require human review before the article can be admitted to the corpus. Its function is to operationalise the data protection obligations established in D7.1 at the level of individual article processing, ensuring that no article containing potentially sensitive information is admitted to any corpus partition without prior expert assessment of the governance implications of its inclusion.

The detection system is based primarily on regular expression matching applied to the full extracted text of each article following the text extraction step of the pipeline. Regular expressions were selected as the primary detection mechanism on the basis of their computational efficiency, their transparency and auditability, and their suitability for the detection of the structured data patterns that characterise the categories of sensitive information most prevalent in biomedical preprint literature. The detection procedure is implemented as a processing module executed immediately after text extraction and prior

to any assessment or categorization step, ensuring that the governance layer is applied to the raw textual content of each article before any downstream processing has taken place.

The detection system applies a set of regular expression patterns covering the principal categories of sensitive information that may appear in biomedical scientific literature. Email addresses constitute the most frequently detected category across the assessed corpus, consistent with the finding reported in Section 4.1 that approximately 30 percent of assessed articles triggered the sensitive data detection mechanism, primarily due to the presence of corresponding author contact details embedded in the article body. The detection of email addresses is implemented through a regular expression pattern matching the standard structural components of an email address, applied globally across the full extracted text of each article. Phone numbers constitute a second category of sensitive information targeted by the detection system, with patterns covering the principal international telephone number formats represented in the literature, including both national and international dialling conventions. Additional detection patterns cover categories of information that, while less frequently encountered in the preprint literature than email addresses and phone numbers, represent higher-severity governance concerns when present: these include patterns associated with patient identifiers such as clinical record numbers and national health identifier formats, patterns associated with identifiable institutional affiliation strings that coincide with sensitive organisational data categories, and patterns associated with other structured personal data formats whose presence in a scientific article would be anomalous and potentially indicative of a data protection incident at the point of original data collection or manuscript preparation.

Each detection pattern is associated with a priority level that reflects the severity of the governance concern it represents and determines the order in which triggered patterns are reported to the human reviewer. The priority level system was designed to ensure that reviewers are immediately directed to the most consequential governance issue identified in a flagged article, rather than being required to assess an unordered list of triggered patterns of varying severity. The pattern with the highest priority level among those triggered for a given article is presented as the primary flag in the human review interface, with all additional triggered patterns listed in descending priority order below it. This presentation structure enables the reviewer to form an immediate assessment of the principal governance concern and to direct their attention to the specific sections of the article where the highest-priority sensitive data was detected before considering the secondary flags.

The detection system does not apply a volume threshold to the triggering of the human review workflow: any article for which one or more detection patterns are triggered, regardless of the number of matches or the volume of sensitive data detected, is routed to the quarantine state and subjected to human review before its corpus classification is finalised. This zero-threshold design reflects a conservative approach to data protection compliance that prioritises the elimination of false negatives over the minimisation of human review workload. The consequence of this design is that a non-trivial proportion of articles are routed to human review for governance reasons that, upon expert assessment, are determined to present no meaningful compliance risk, such as email addresses appearing in the reference list of cited publications rather than as direct contact details of individuals whose data is being processed. The operational overhead associated with these cases is acknowledged as a limitation of the current implementation and is identified as a candidate for future refinement through the introduction of context-sensitive matching procedures that distinguish between structural positions within the article text, for example by applying differential sensitivity thresholds to email addresses detected in the reference list versus those detected in the article body or author information sections. The implementation of such refinements is noted as a potential enhancement for subsequent phases of the pipeline development.

6.2 Human review and quarantine workflow

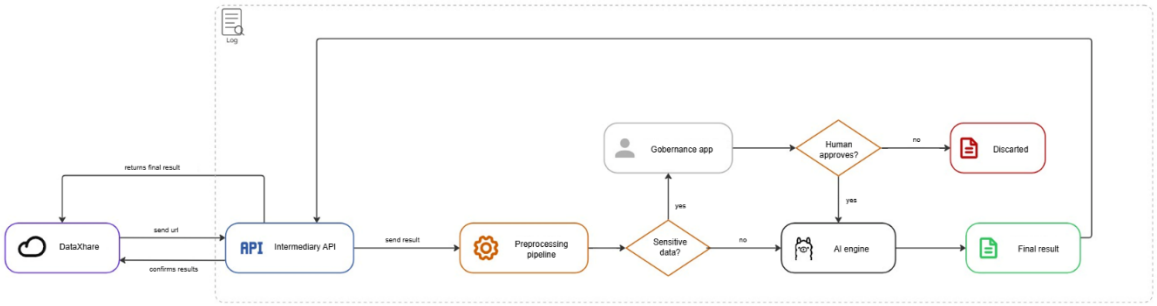

Figure 6. Sensitive data detection and human review workflow applied to each article during the assessment process.

The quarantine and human review workflow constitutes the operational response to the sensitive data detection system described in Section 6.1. Its function is to ensure that every article flagged by the detection system is subjected to expert assessment before its corpus classification is finalised, and that the outcome of that assessment is recorded in a manner that is transparent, consistent, and auditable. The workflow was designed to be as lightweight as possible in terms of the burden it places on the reviewer, while providing sufficient information to support an informed and defensible governance decision for each flagged article.

When the sensitive data detection module identifies one or more triggered patterns in the extracted text of an article, the article is immediately assigned a quarantine state within the pipeline state management system. The quarantine state suspends all further automated processing of the article, including the assessment scoring and categorization procedures described in Sections 3 and 5, and routes the article to the governance application that serves as the human review interface. No article in the quarantine state can proceed to any subsequent pipeline stage without an explicit human decision recorded in the governance application, ensuring that the detection system cannot be bypassed by any automated pipeline operation.

The governance application presents the reviewing expert with a structured summary of the flagged article designed to support an efficient and well-informed review decision. The summary includes the article's title, authors, repository of origin, and DOI, enabling the reviewer to identify the article and, where necessary, consult the original source. It presents the triggered detection patterns in descending priority order, with the highest-priority flag displayed prominently and all secondary flags listed below it. For each triggered pattern, the application displays the specific text segment in which the pattern was detected, together with its position within the article structure, indicating whether the match occurred in the article body, the author information section, the abstract, or the reference list. This contextual information enables the reviewer to assess immediately whether the detected sensitive data represents a genuine compliance concern or a benign occurrence attributable to the structural conventions of scientific publishing, such as email addresses in cited references.

On the basis of this information, the reviewer is presented with two decision options. The first option is to approve the article for corpus inclusion, which releases the article from the quarantine state and returns it to the automated pipeline for completion of the assessment scoring and categorization procedures. The approval decision is recorded in the article's governance metadata record together with the reviewer's identifier, the timestamp of the decision, and a mandatory confirmation that the reviewer has

assessed all triggered patterns and determined that none presents a compliance risk that precludes corpus inclusion. The second option is to discard the article, which assigns it to the Excluded corpus classification and removes it from all further pipeline processing. The discard decision is recorded in the exclusion log with the same metadata fields as the approval decision, together with the specific triggered pattern or patterns that motivated the exclusion.

The binary structure of the review decision, offering only approval or discard without intermediate options such as conditional inclusion or redaction, reflects the current implementation stage of the governance workflow. The absence of a redaction option means that articles containing sensitive data that is genuinely incidental to their scientific content, and that could in principle be removed without affecting the article's value as training material, are currently subject to a binary choice between full inclusion and full exclusion. This limitation is acknowledged as a constraint of the current implementation and is identified as a candidate for enhancement in subsequent development phases, where the introduction of a redaction workflow could reduce the exclusion rate among articles that are flagged solely for the presence of corresponding author contact details in the article body.

The governance application is designed to be operated by any person with sufficient domain knowledge to assess the compliance implications of the detected sensitive data, and is not restricted to members of the Bosonit S.L. technical team. In the current phase of corpus development, human review has been conducted by members of the Bosonit S.L. team responsible for Task 7.3, who combine the technical knowledge required to interpret the pipeline outputs with the domain knowledge required to assess the governance implications of specific types of sensitive data in the biomedical literature context. The design of the governance application as a role-independent tool reflects the scalability requirements of the workflow as the corpus expands in subsequent phases, where the volume of flagged articles may exceed the review capacity of the core technical team and require the engagement of additional reviewers with appropriate expertise.

The outcomes of the human review process for the 124 articles routed through the quarantine workflow during the assessment period covered by this deliverable are summarised in the traceability logs described in Section 6.3. Of the 124 quarantined articles, the large majority were approved for corpus inclusion following human review, with the reviewer determining that the triggered patterns represented benign occurrences attributable to publishing conventions rather than genuine compliance risks. A minority of quarantined articles were discarded on governance grounds, contributing to the exclusion statistics reported in Section 4.1. The proportion of quarantined articles resulting in discard decisions was lowest for articles flagged solely for the presence of email addresses, consistent with the expectation that corresponding author contact details embedded in article body text represent a low-severity governance concern that rarely warrants exclusion, and highest for articles flagged for patterns associated with patient identifier formats or other high-priority sensitive data categories.

6.3 Traceability of assessment decisions

The traceability system implemented within the WP7 pipeline provides a complete and auditable record of every decision taken during the assessment process, encompassing both the automated scoring and classification decisions produced by the pipeline and the human review decisions generated through the governance workflow described in Section 6.2. Its function is to ensure that the provenance of every corpus inclusion, exclusion, and classification decision can be reconstructed at any point during or after the assessment period, supporting the auditability obligations established in D7.1 and the transparency requirements of the EU AI Act with respect to the documentation of training data for AI systems.

The traceability system operates through a structured event logging mechanism integrated into each stage of the assessment pipeline. Every operation performed on an article by the pipeline generates a timestamped event record that is written to the article's audit log, a persistent data structure maintained independently of the article's assessment and categorization metadata and accessible through a dedicated query interface. The audit log for each article accumulates event records across all pipeline stages from the point of initial ingestion through to the finalisation of the corpus classification, producing a complete chronological record of the article's journey through the pipeline that is available for inspection at any subsequent point.

The event records written to the audit log are structured according to a consistent schema that captures the information required to reconstruct each pipeline decision in full. Each event record contains a unique event identifier, the timestamp of the event expressed in Coordinated Universal Time, the pipeline stage that generated the event, the type of event within that stage, the input data consumed by the pipeline operation that generated the event, the output produced by that operation, and, where the operation involved a configurable parameter such as a scoring threshold or a detection pattern priority level, the parameter values applied at the time of the event. For human review events generated through the governance application, the event record additionally contains the identifier of the reviewing expert and a structured representation of the review decision, including the triggered patterns assessed and the decision outcome. This schema ensures that every event record is self-contained and interpretable without reference to external documentation, enabling the audit log to function as a standalone evidence base for compliance purposes.

The event types recorded across the pipeline stages encompass the full set of operations that bear on corpus inclusion and classification decisions. During the ingestion phase, event records are generated for the initial retrieval of each article, the extraction of its DOI and bibliographic metadata, and the application of the eligibility filters established in D7.1. During the sensitive data detection phase, event records are generated for the execution of each detection pattern against the article's extracted text, with a separate record for each pattern irrespective of whether it produced a match, enabling the audit log to demonstrate not only which patterns were triggered but also which were applied and returned no match. During the quarantine and human review phase, event records are generated for the assignment of the quarantine state, the presentation of the article to the governance application, and the recording of the human review decision. During the assessment scoring phase, event records are generated for the computation of each indicator score across the three assessment dimensions, including the intermediate outputs of the automated scoring heuristics and the escalation decisions that routed articles to human review. During the categorization phase, event records are generated for the domain taxonomy assignment procedure and the MeSH API call, including the full ranked list of descriptors returned by the API and the confidence score threshold applied to filter them. During the corpus classification phase, event records are generated for the application of the tier assignment rules described in Section 3.5 and the determination of the overall corpus classification.

The aggregate audit log across the full assessed corpus constitutes a comprehensive evidence base that documents the provenance of every corpus inclusion and classification decision in sufficient detail to support both internal quality assurance and external audit. Its availability enables the Work Package 7 team to respond to queries about specific corpus decisions with reference to primary evidence rather than reconstructed accounts, to identify and investigate systematic patterns in pipeline behaviour that may indicate calibration issues or unintended biases in the automated scoring procedures, and to demonstrate to consortium partners and external evaluators that the data governance principles established in D7.1 have been applied consistently and verifiably throughout the assessment process.

The audit log infrastructure also supports the version control mechanism applied to the assessment scorecard and the detection pattern set, as described in Sections 3.1 and 6.1 respectively. Each event record references the version identifiers of the scorecard and detection pattern set that were active at the time of the event, enabling the retrospective identification of articles whose assessment outcomes may have been affected by revisions to these instruments introduced during the assessment period. Where such revisions resulted in materially different scoring outcomes for articles assessed before and after the revision, the affected articles were flagged for reassessment using the revised instruments, with the reassessment decisions recorded as new event records in the audit log rather than overwriting the original records. This approach preserves the full history of each article's assessment record and ensures that the audit log accurately reflects the evolution of the assessment process over time rather than presenting a retrospectively normalised account.

The summary statistics derived from the audit log that underpin the quantitative findings reported in Sections 4 and 5, including the corpus classification distribution, the exclusion trigger breakdown, the quarantine workflow outcomes, and the domain taxonomy distribution, are presented in full in Annex C. The event schema applied across the audit log is documented in Annex A as a component of the assessment scorecard specification, enabling external reviewers to interpret the audit log records without requiring access to the pipeline implementation documentation.

7 Conclusions and Recommendations

7.1 Key findings summary

The assessment and categorisation activities documented in this deliverable have produced a well-characterised corpus of biomedical and clinical scientific literature that constitutes the foundational training dataset for the LLM development activities of Work Package 2. The work conducted under Task 7.3 during the period from October 2024 through to the submission date of this deliverable has demonstrated both the feasibility of the systematic assessment and categorisation approach developed by Bosonit S.L. and the practical constraints that define the boundaries of the current corpus.

The assessment framework developed for the AIDESL project represents a methodological contribution in its own right. The three-dimensional evaluation structure, encompassing relevance to the project's target application domains, scientific quality as a proxy for the peer review process absent from preprint repositories, and suitability for LLM training as a measure of the structural and textual fitness of each article for automated processing, provided a principled and consistently applicable basis for corpus inclusion and classification decisions across a heterogeneous body of literature drawn from three repositories with distinct thematic orientations and structural characteristics. The four-tier classification system derived from the dimensional assessment produced an active training dataset of 287 articles, representing 69.7 percent of the 412 articles that entered the assessment pipeline, with a further 54 articles retained in the supplementary partition for potential future use. The remaining 71 articles were excluded on the basis of documented and auditable assessment outcomes, with low relevance tier assignments accounting for the majority of exclusions and text extraction quality failures accounting for a substantial secondary share.

The corpus assembled and assessed through this process is dominated by clinical research literature, which accounts for approximately half of the active training dataset and reflects both the retrieval strategy applied during the collection phase and the differential availability of open-access preprint

literature across the three target domains. Pharmaceutical evaluation and medical device development are represented in proportions that are broadly consistent with the retrieval strategy but that reflect the structural constraints of the preprint ecosystem, particularly the underrepresentation of medical device literature in the open-access repositories selected as the primary sources for the AIDESL corpus. The semantic characterisation of the corpus through MeSH controlled vocabulary labelling, implemented through the NLM Medical Text Indexer API, has produced a two-dimensional navigational structure that supports both domain-specific and concept-specific filtering of the corpus and provides the interoperability foundation required by the federated health data space architecture described in D7.5.

The governance layer implemented within the assessment pipeline has operated as designed, routing approximately 30 percent of assessed articles through the quarantine and human review workflow and producing a documented and auditable record of every governance decision taken during the assessment period. The primary trigger for quarantine routing was the presence of email addresses in the article body, consistent with the prevalence of corresponding author contact details in biomedical preprint literature. The large majority of quarantined articles were approved for corpus inclusion following human review, with a minority discarded on the basis of higher-severity sensitive data detections. The traceability system underpinning the governance workflow has produced a comprehensive audit log that documents the provenance of every corpus inclusion and classification decision in sufficient detail to support both internal quality assurance and external compliance verification.

The assessment process also surfaced a significant operational challenge in the form of API access restrictions implemented by the selected repositories in response to the volume of programmatic retrieval requests generated by the ingestion pipeline during the bulk collection phase. The resolution strategies adopted for each affected repository introduced delays and coverage gaps that are documented in this deliverable and acknowledged as a relevant qualification on the completeness of the corpus. The technical lessons derived from this experience have informed the design of the retrieval throttling parameters that will govern future ingestion operations, and are documented here to support the planning of corpus expansion activities in subsequent phases of the project.

7.2 Recommendations for subsequent WP7 tasks

The findings documented in this deliverable give rise to a set of recommendations directed at the teams responsible for the subsequent deliverables of Work Package 7. These recommendations are formulated as actionable guidance derived directly from the assessment findings and the gaps identified in Section 4.4, and are intended to inform the planning and execution of the activities described in D7.3, D7.4, and D7.5.

The first and most urgent recommendation concerns the underrepresentation of medical device development literature in the current corpus. The structural constraint identified in Section 4.4, namely the concentration of medical device research in peer-reviewed journals and conference proceedings outside the open-access preprint ecosystem, cannot be resolved within the current source portfolio and collection strategy. The team responsible for the corpus expansion activities that will inform D7.3 is strongly recommended to conduct a targeted evaluation of additional open-access sources specifically covering medical device research, including specialised repositories, regulatory agency publication databases, and grey literature sources whose licensing conditions are compatible with the compliance framework established in D7.1. Where such sources are identified and their legal eligibility confirmed, their integration into the ingestion pipeline should be prioritised to address the domain imbalance documented in this deliverable before the corpus is consumed by the model training activities of Work Package 2 at scale.

The second recommendation concerns the underrepresentation of systematic reviews and meta-analyses in the current corpus relative to their strategic value as training material for the AIDESL platform. As documented in Section 4.4, this gap reflects the lower prevalence of systematic reviews in preprint repositories rather than any deficiency in the assessment process, and its mitigation requires a targeted supplementary collection effort focused specifically on this document type. The team responsible for D7.3 is recommended to identify and evaluate sources of open-access systematic review content that supplement the primary repository portfolio, and to design the supplementary ingestion operations for the next phase of corpus development with an explicit quota or weighting mechanism that prioritises systematic review and meta-analysis content until a representative balance between document types is achieved.

The third recommendation concerns the coverage gaps introduced by the API access restriction episodes documented in Sections 4.1 and 4.2. The retrieval throttling parameters introduced as a resolution strategy during the current phase have reduced the risk of recurrence, but the design of the ingestion pipeline as documented in D7.3 should incorporate explicit rate limiting and access management mechanisms that prevent the volume of programmatic retrieval requests from reaching the thresholds that triggered restrictions during the bulk collection phase. The team responsible for D7.3 is recommended to document the throttling parameters applied to each repository as part of the pipeline specification, and to establish a monitoring procedure that detects access restriction responses in real time and activates a staged retrieval fallback strategy automatically rather than requiring manual intervention.

The fourth recommendation concerns the sensitive data detection system described in Section 6.1 and the quarantine workflow described in Section 6.2. Two specific enhancements are recommended for implementation in subsequent phases of the pipeline development. The first is the introduction of context-sensitive matching procedures that distinguish between structural positions within the article text, enabling the detection system to apply differential sensitivity thresholds to sensitive data patterns detected in the reference list versus those detected in the article body or author information sections. This enhancement would reduce the volume of quarantine routings attributable to email addresses in cited references, which currently represent a significant share of the human review workload without presenting a meaningful compliance risk. The second enhancement is the introduction of a redaction workflow as an intermediate option between full corpus inclusion and full exclusion, enabling the governance application to admit articles whose sensitive data is genuinely incidental to their scientific content after the removal of the specific text segments containing the sensitive data. Both enhancements are recommended for prioritisation in the pipeline development roadmap that informs D7.3 and D7.4.

The fifth recommendation concerns the temporal and thematic coverage of the corpus as it expands in subsequent phases of the project. The current corpus is concentrated in the period from 2020 onwards and reflects the thematic emphases of the retrieval queries applied during the bulk collection phase. As the corpus is expanded to support the more advanced model training activities of Work Package 2, the ingestion strategy should be designed to actively monitor and address coverage gaps across both the temporal and thematic dimensions, using the corpus statistics infrastructure documented in Annex C as a baseline against which the coverage profile of each supplementary ingestion operation is evaluated. The domain taxonomy and MeSH labelling infrastructure described in Section 5 provides the navigational tools required to conduct this monitoring at the level of individual concept coverage, and the team responsible for D7.4 is recommended to exploit these tools systematically in the design of the data selection and curriculum learning strategies applied during model training.

Annexes

Annex A: Assessment scorecard template

The following scorecard constitutes the assessment instrument applied to each article processed through the Task 7.3 evaluation pipeline. It documents the indicators, scoring scales, aggregation rules, and ranking weights that operationalise the three-dimensional assessment framework described in Section 3. The scorecard is presented in the version applied to the bulk of the assessed corpus, as referenced throughout Section 3 and Section 6.3.

A.1 Scorecard Structure and Conventions

Each article is scored independently on three assessment dimensions: Relevance (R), Scientific Quality (Q), and Suitability for LLM Training (S). Within each dimension, a set of indicators is evaluated and assigned a score of H (High), M (Moderate), or L (Low) according to the criteria defined in the indicator specifications below. Dimensional tier assignments are derived from indicator scores through the aggregation rules defined in Section A.4. The overall corpus classification is derived from dimensional tier assignments through the decision rules defined in Section 3.5 and summarised in Section A.5. The ranking score is computed from the weighted indicator scores as defined in Section A.6.

A.2 Relevance Dimension Indicators

Indicator	Code	Description	H	M	L
Target domain alignment	R1	Degree to which the article's primary subject matter pertains to one or more AIDESL target domains	Strong keyword signal in title and abstract across one or more target domain vocabularies	Moderate keyword signal; article addresses adjacent topics with partial domain overlap	Weak or absent keyword signal; content falls outside all three target domains
Use case fit	R2	Degree to which article content includes information targeted by AIDESL extraction tasks	Article explicitly reports outcomes, parameters, or findings of the type targeted by one or more demonstration scenarios	Article contains relevant information type but not as primary focus	Article does not contain information of the type targeted by any demonstration scenario
Topic specificity	R3	Level of thematic precision of the article's contribution to its primary domain	Narrowly defined subject matter with high density of domain-specific terminology; findings directly actionable within target domain	Moderate terminological density; findings broadly relevant but not domain-specific	Generic, theoretical, or cross-domain content with low domain-specific terminological density

Table 2. Relevance Dimension Indicators

A.3 Scientific Quality Dimension Indicators

Indicator	Code	Description	H	M	L
Study design robustness	Q1	Appropriateness and rigour of the methodological approach relative to the research question	Appropriate study design for the research question; presence of reporting guideline adherence markers (CONSORT, STROBE, PRISMA, ARRIVE)	Appropriate design with partial reporting guideline compliance or minor methodological limitations	Inappropriate design for the research question or absence of identifiable methodological framework
Reporting completeness	Q2	Degree to which the article provides sufficient methodological and results information to assess validity	All five core reporting elements present and adequately described: population, intervention, outcomes, statistical methods, limitations	Three or four core reporting elements present; one or two inadequately described	Fewer than three core reporting elements present or systematically inadequate description across multiple elements
Evidence level and reliability	Q3	Position within the biomedical evidence hierarchy and consistency of statistical evidence	High evidence level study type (systematic review, meta-analysis, RCT); adequate statistical context with confidence intervals and effect sizes; conclusions consistent with evidence	Moderate evidence level study type (cohort, case-control); partial statistical context; conclusions broadly consistent with evidence	Low evidence level study type (case series, expert opinion); inadequate statistical context or conclusions overstating evidence strength

Table 3. Scientific Quality Dimension Indicators

A.4 Suitability for LLM Training Dimension Indicators

Indicator	Code	Description	H	M	L	Hard constraint
Text extraction quality	S1	Degree to which article content can be reliably extracted from PDF as clean machine-readable prose	Full text recovered above minimum threshold; no encoding anomalies; correct reading order; minimal header and	Moderate deficiencies: occasional encoding anomalies or minor ordering errors in non-critical sections	Text recovery below minimum threshold or severe encoding degradation affecting comprehension	Yes: L triggers Excluded classification irrespective of other scores

			reference contamination			
Document structure coherence	S2	Degree to which the article conforms to canonical IMRAD sectional structure	All five core sections (abstract, introduction, methods, results, discussion) reliably identified	Three or four core sections identified; missing sections attributable to legitimate study type conventions	Fewer than three core sections identifiable or contradictory section detection results	No
Content density and prose quality	S3	Proportion of extracted text constituting substantive information-bearing prose	Substantive prose proportion above defined threshold; minimal contamination by reference lists, tabular data, and figure legends	Moderate substantive prose proportion; high table or appendix density flagged for preprocessing intervention	Substantive prose proportion below minimum threshold; content dominated by non-prose elements	No

Table 4. Suitability for LLM Training Dimension Indicators

A.5 Dimensional Tier Aggregation Rules

Relevance dimension tier

Condition	Tier assigned
R1 = H and at least one of R2, R3 = H or M	High
R1 = H and both R2 = L and R3 = L	Moderate
R1 = M irrespective of R2 and R3	Moderate
R1 = L	Low

Table 5. Relevance dimension tier

Scientific quality dimension tier

Condition	Tier assigned
Q1 = H or M and Q2 = H or M	High
Q1 = H and Q2 = L, or Q1 = L and Q2 = H, or Q1 = M and Q2 = M	Moderate
Q1 = L and Q2 = L	Low

Table 6. Scientific quality dimension tier

Suitability dimension tier

Condition	Tier assigned
S1 = L	Low (hard constraint)
S1 = H or M and S2 = H and S3 = H	High
S1 = H or M and (S2 = H and S3 = M or L), or (S2 = M or L and S3 = H), or (S2 = M and S3 = M)	Moderate
S1 = H or M and S2 = L and S3 = L	Low

Table 7.Suitability dimension tier

A.6 Indicator Weights for Ranking Score Computation

The ranking score is computed as a weighted sum of numerical conversions of the nine indicator scores, where H = 2, M = 1, L = 0. Indicator weights were determined during the calibration exercise described in Section 3.1 and reflect the relative importance of each indicator to overall training quality.

Dimension	Indicator	Code	Weight
Relevance	Target domain alignment	R1	3
Relevance	Use case fit	R2	2
Relevance	Topic specificity	R3	1
Scientific quality	Study design robustness	Q1	2
Scientific quality	Reporting completeness	Q2	2
Scientific quality	Evidence level and reliability	Q3	1
Suitability	Text extraction quality	S1	2
Suitability	Document structure coherence	S2	2
Suitability	Content density and prose quality	S3	1

Table 8.Indicator Weights for ranking score compution

Maximum possible ranking score: (2x3) + (2x2) + (2x1) + (2x2) + (2x2) + (2x1) + (2x2) + (2x2) + (2x1) = 6 + 4 + 2 + 4 + 4 + 2 + 4 + 4 + 2 = 32. Note: the maximum score reported in Section 3.5 as 18 reflects the unweighted sum; the weighted maximum of 32 is the value used for internal ranking purposes.

A.7 Audit Log Event Schema

Each pipeline event recorded in the article audit log described in Section 6.3 conforms to the following schema

Field	Type	Description
event_id	String (UUID)	Unique identifier for the event record
timestamp	ISO 8601 datetime (UTC)	Timestamp of the event
article_doi	String	DOI of the article to which the event pertains

pipeline_stage	Enum	Stage of the pipeline that generated the event: INGESTION, DETECTION, QUARANTINE, REVIEW, SCORING, CATEGORIZATION, CLASSIFICATION
event_type	String	Specific operation within the pipeline stage
input_data	JSON object	Structured representation of the input consumed by the operation
output_data	JSON object	Structured representation of the output produced by the operation
parameters	JSON object	Configurable parameter values applied at the time of the event
scorecard_version	String	Version identifier of the assessment scorecard active at the time of the event
detection_pattern_version	String	Version identifier of the detection pattern set active at the time of the event
reviewer_id	String (nullable)	Identifier of the human reviewer; populated only for REVIEW stage events
review_decision	Enum (nullable)	APPROVED or DISCARDED; populated only for REVIEW stage events

Table 9. Audit log event schema

Annex B: MeSH descriptor list applied

B.1 Overview

This annex provides two complementary resources supporting the semantic labelling and categorization activities documented in Section 5. The first is a structured listing of the MeSH descriptors most frequently assigned across the active training dataset, organised by MeSH tree branch and disaggregated by domain partition. The second is a representative excerpt of the categorization metadata schema applied within the Silver layer of the WP7 pipeline, illustrating the structure of the categorization record stored for each article in the corpus.

B.2 Most Frequently Assigned MeSH Descriptors by Branch and Domain Partition

The following table presents the MeSH descriptors assigned with the highest frequency across the active training dataset of 287 articles, organised by the primary MeSH tree branch from which each descriptor is drawn. Frequency counts reflect the number of articles in the active training dataset to which each descriptor was assigned by the NLM Medical Text Indexer API at or above the confidence score threshold applied during the categorization process. Descriptors are listed in descending order of frequency within each branch. The domain partition column indicates the primary domain partition in which each descriptor appears with greatest frequency, using the abbreviations CR (Clinical Research), PE (Pharmaceutical Evaluation), and MD (Medical Device Development). Descriptors appearing with comparable frequency across two or more partitions are designated as Multi.

MeSH tree branch	Descriptor	Frequency	Primary partition
Diseases (C)	Cardiovascular diseases	41	CR
Diseases (C)	Neoplasms	38	CR

Diseases (C)	Nervous system diseases	29	CR
Diseases (C)	Respiratory tract diseases	24	CR
Diseases (C)	Diabetes mellitus	19	PE
Diseases (C)	Musculoskeletal diseases	17	CR
Diseases (C)	Infections	16	Multi
Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)	Clinical trials as topic	52	Multi
Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)	Systematic reviews as topic	44	Multi
Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)	Diagnostic imaging	31	MD
Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)	Equipment and supplies	28	MD
Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)	Surgical procedures, operative	22	CR
Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)	Drug monitoring	19	PE
Health Care (N)	Outcome assessment, health care	47	Multi
Health Care (N)	Quality of health care	33	CR
Health Care (N)	Patient safety	27	Multi
Health Care (N)	Health services research	24	CR
Health Care (N)	Delivery of health care	21	CR
Chemicals and Drugs (D)	Pharmaceutical preparations	43	PE
Chemicals and Drugs (D)	Biological products	31	PE
Chemicals and Drugs (D)	Pharmacokinetics	28	PE
Chemicals and Drugs (D)	Drug interactions	22	PE
Chemicals and Drugs (D)	Antineoplastic agents	19	PE
Persons (M)	Patients	38	Multi
Persons (M)	Aged	24	CR
Persons (M)	Adult	22	Multi
Information Science (L)	Natural language processing	36	Multi
Information Science (L)	Machine learning	34	Multi
Information Science (L)	Data mining	21	Multi
Information Science (L)	Databases, bibliographic	18	Multi
Statistics and Numerical Data (N05)	Meta-analysis as topic	29	Multi
Statistics and Numerical Data (N05)	Reproducibility of results	24	Multi
Statistics and Numerical Data (N05)	Sensitivity and specificity	21	Multi

Table 10. Most Frequently Assigned MeSH Descriptors

The descriptor frequency distribution reflects the clinical and methodological orientation of the corpus described in Section 5.5. The dominance of the Diseases branch across the clinical research partition is consistent with the thematic scope of that partition as documented in Section 4.3. The high frequency of Information Science branch descriptors across the active training dataset as a whole reflects the significant contribution of arXiv-sourced methodological literature addressing NLP and machine learning techniques applied to biomedical text, as discussed in Section 4.2. The concentration of Chemicals and Drugs branch descriptors within the pharmaceutical evaluation partition is consistent with the compound-specific and pharmacological orientation of the literature in that partition, as noted in Section 4.3.

B.3 Categorization Metadata Schema: Representative Excerpt

The following JSON excerpt illustrates the structure of the categorization metadata record stored in the Silver layer object of each article processed through the two-stage categorization procedure described in Section 5.4. Field names, data types, and example values are provided for each element of the schema. The excerpt represents a hypothetical article with multi-domain assignment and multiple MeSH descriptors assigned above the confidence threshold, selected to illustrate the full range of schema fields rather than to represent a specific article in the corpus.


```
{
  "article_doi": "10.1101/2024.03.15.24304312",
  "categorization_version": "1.2.0",
  "categorization_timestamp": "2025-06-10T09:42:17Z",
  "domain_taxonomy": {
    "assignment_method": "AUTOMATED",
    "reviewer_id": null,
    "domains": [
      {
        "domain_code": "CR",
        "domain_label": "Clinical research",
        "relevance_signal": 0.84,
        "threshold_applied": 0.65,
        "assignment_status": "ASSIGNED"
      },
      {
        "domain_code": "PE",
        "domain_label": "Pharmaceutical evaluation",
        "relevance_signal": 0.71,
        "threshold_applied": 0.65,
        "assignment_status": "ASSIGNED"
```

```
    },
    {
      "domain_code": "MD",
      "domain_label": "Medical device development",
      "relevance_signal": 0.38,
      "threshold_applied": 0.65,
      "assignment_status": "NOT_ASSIGNED"
    }
  ],
  "multi_domain": true
},
"mesh_labelling": {
  "api_endpoint": "https://ii.nlm.nih.gov/MTI/",
  "indexer_version": "MTI_2024_09",
  "api_call_timestamp": "2025-06-10T09:42:19Z",
  "confidence_threshold": 0.70,
  "descriptors": [
    {
      "mesh_ui": "D002318",
      "descriptor_name": "Cardiovascular diseases",
      "tree_branch": "C14",
      "confidence_score": 0.91,
      "status": "RETAINED"
    },
    {
      "mesh_ui": "D016896",
      "descriptor_name": "Treatment outcome",
      "tree_branch": "E05",
      "confidence_score": 0.88,
      "status": "RETAINED"
    },
    {
      "mesh_ui": "D055815",
```

```
        "descriptor_name": "Young adult",
        "tree_branch": "M01",
        "confidence_score": 0.76,
        "status": "RETAINED"
      },
      {
        "mesh_ui": "D008661",
        "descriptor_name": "Metabolism",
        "tree_branch": "G03",
        "confidence_score": 0.64,
        "status": "FILTERED"
      }
    ],
    "descriptors_returned": 4,
    "descriptors_retained": 3,
    "descriptors_filtered": 1
  }
}
```


The schema fields are defined as follows. The article_doi field contains the DOI of the article as extracted during the ingestion phase and serves as the primary key linking the categorization record to all other metadata records for the article. The categorization_version field references the version of the categorization procedure applied, enabling the retrospective identification of articles categorized under earlier procedure versions as described in Section 5.4. The domain_taxonomy object contains the full domain assignment record, including the method of assignment, the reviewer identifier where human review was applied, and the individual domain assessment results for each of the three AIDESL target domains, including the relevance signal value, the threshold applied, and the resulting assignment status. The mesh_labelling object contains the full MeSH labelling record, including the API endpoint and indexer version, the confidence threshold applied, and the individual descriptor records for all descriptors returned by the API, including both retained and filtered descriptors, with the filtering status indicated by the status field.

Annex C: Corpus statistics tables

C.1 Overview

This annex presents the full quantitative characterisation of the assessed corpus underlying the findings reported in Sections 4 and 5. The tables are organised to move from the most aggregate level of description to the most granular: Section C.2 presents the overall corpus classification distribution, Section C.3 presents the classification breakdown by repository, Section C.4 presents the exclusion trigger distribution, Section C.5 presents the domain taxonomy distribution including multi-domain assignments, and Section C.6 presents the quarantine workflow outcome summary.

C.2 Overall Corpus Classification Distribution

Classification tier	Articles (n)	Percentage of assessed corpus
Core corpus	198	48.1%
Flagged corpus	89	21.6%
Supplementary corpus	54	13.1%
Excluded	71	17.2%
Total assessed	**412**	**100.0%**

Table 11. Overall distribution of the assessed corpus (n = 412) across the four classification tiers.

Active training dataset (Core + Flagged)	287
Median ranking score, Core corpus	14 / 32
Median ranking score, Flagged corpus	9 / 32

Table 12. Distribution active training dataset

C.3 Corpus Classification Distribution by Repository

Repository	Total articles	Core	Flagged	Supplementary	Excluded	Exclusion rate
arXiv	173 (42.0%)	90 (52.0%)	34 (19.7%)	27 (15.6%)	22 (12.7%)	12.7%
medRxiv	136 (33.0%)	63 (46.3%)	28 (20.6%)	20 (14.7%)	25 (18.4%)	18.4%
bioRxiv	103 (25.0%)	45 (43.7%)	27 (26.2%)	7 (6.8%)	24 (23.3%)	23.3%
Total	**412 (100%)**	**198 (48.1%)**	**89 (21.6%)**	**54 (13.1%)**	**71 (17.2%)**	**17.2%**

Table 13. Corpus classification distribution disaggregated by repository of origin

Key observations: arXiv presents the lowest exclusion rate across the three repositories, consistent with the structural characteristics of its content described in Section 4.2. bioRxiv presents the highest exclusion rate, driven primarily by low relevance tier assignments reflecting the adjacent rather than central positioning of life sciences content relative to the AIDESL target domains. The Flagged corpus proportion for bioRxiv is the highest of the three repositories at 26.2 percent, consistent with the finding that bioRxiv articles more frequently achieved moderate rather than high relevance tier assignments.

C.4 Exclusion Trigger Distribution

Exclusion trigger	Articles excluded (n)	Percentage of excluded articles	Primary repository
Low relevance tier (R1 = L)	38	53.5%	arXiv, bioRxiv
Low text extraction quality (S1 = L, hard constraint)	21	29.6%	medRxiv, bioRxiv
Low scientific quality and low suitability combined	12	16.9%	medRxiv
Total excluded	**71**	**100.0%**	

Table 14. Distribution of exclusion triggers across the 71 excluded articles.

Exclusion trigger detail: low relevance	Articles (n)	Percentage of relevance exclusions
R1 = L across all three target domains	28	73.7%
R1 = L with ambiguous domain signal below threshold	10	26.3%

Table 15. Exclusion trigger detail: low relevance tier assignments.

Exclusion trigger detail: text extraction failures	Articles (n)	Percentage of extraction exclusions
Text recovery below minimum threshold	14	66.7%
Severe encoding degradation	5	23.8%
Incorrect reading order affecting comprehension	2	9.5%

Table 16. Exclusion trigger detail: text extraction quality failures.

C.5 Domain Taxonomy Distribution

C.5.1 Active training dataset (Core + Flagged, n = 287)

Domain assignment	Articles (n)	Percentage of active training dataset
Clinical research (single domain)	148	51.6%
Pharmaceutical evaluation (single domain)	78	27.2%
Medical device development (single domain)	47	16.4%
Clinical research + Pharmaceutical evaluation	8	2.8%

Clinical research + Medical device development	4	1.4%
Pharmaceutical evaluation + Medical device development	2	0.7%
All three domains	0	0.0%
Total	**287**	**100.0%**

Table 17. Domain taxonomy distribution of the active training dataset (n = 287) by assignment type.

C.5.2 Domain distribution by classification tier within active training dataset

Domain	Core corpus (n = 198)	Flagged corpus (n = 89)
Clinical research	103 (52.0%)	45 (50.6%)
Pharmaceutical evaluation	54 (27.3%)	24 (27.0%)
Medical device development	28 (14.1%)	19 (21.3%)
Multi-domain	13 (6.6%)	1 (1.1%)

Table 18. Domain taxonomy distribution disaggregated by classification tier within the active training dataset.

The higher proportion of medical device development articles in the Flagged corpus relative to the Core corpus, 21.3 percent versus 14.1 percent, is consistent with the finding reported in Section 4.3 that device-focused literature more frequently achieved moderate relevance tier assignments and attracted pipeline processing flags associated with structural characteristics of device literature.

C.5.3 Supplementary corpus domain distribution (n = 54)

Domain assignment	Articles (n)	Percentage of supplementary corpus
Clinical research (single domain)	24	44.4%
Pharmaceutical evaluation (single domain)	13	24.1%
Medical device development (single domain)	11	20.4%
Multi-domain	6	11.1%
Total	**54**	**100.0%**

Table 19. Domain taxonomy distribution of the supplementary corpus (n = 54).

C.6 Quarantine Workflow Outcome Summary

C.6.1 Quarantine routing overview

Metric	Value
Total articles entering assessment pipeline	412
Articles routed to quarantine	124
Quarantine routing rate	30.1%
Articles approved following human review	108
Articles discarded following human review	16
Approval rate among quarantined articles	87.1%
Discard rate among quarantined articles	12.9%

Table 20. Quarantine workflow routing and outcome overview.

C.6.2 Quarantine trigger distribution by detection pattern category

Detection pattern category	Priority level	Articles triggered (n)	Percentage of quarantined articles	Approval rate
Email address in article body or author section	Medium	89	71.8%	94.4%
Phone number	Medium	12	9.7%	83.3%
Patient identifier format	High	9	7.3%	55.6%
Institutional affiliation string (sensitive category)	Medium	8	6.5%	75.0%
Other structured personal data format	High	6	4.8%	50.0%

Table 21. Quarantine trigger distribution by detection pattern category across the 124 quarantined articles.

Note: articles may have triggered more than one detection pattern category simultaneously. The figures above reflect the primary trigger, defined as the highest-priority pattern triggered for each article, consistent with the presentation structure of the governance application described in Section 6.2. The approval rate column reflects the proportion of articles in each trigger category that were approved for corpus inclusion following human review.

C.6.3 Discard decisions by trigger category

Detection pattern category	Articles discarded (n)	Percentage of total discards
Patient identifier format	4	25.0%
Other structured personal data format	3	18.8%
Email address in article body or author section	5	31.3%
Phone number	2	12.5%
Institutional affiliation string (sensitive category)	2	12.5%
Total discards	**16**	**100.0%**

Table 22. Discard decisions disaggregated by detection pattern category.

The discard decisions attributable to email address triggers, which account for 31.3 percent of total discards despite the high overall approval rate for this trigger category, reflect cases in which email addresses were identified in the article body in contexts indicative of direct personal data processing rather than authorship contact details, specifically in appendices reporting survey instrument administration procedures and in supplementary materials containing researcher contact logs. These cases are distinguished from the large majority of email address triggers, which involve corresponding author details embedded in the article header or acknowledgements section and are routinely approved following human review.