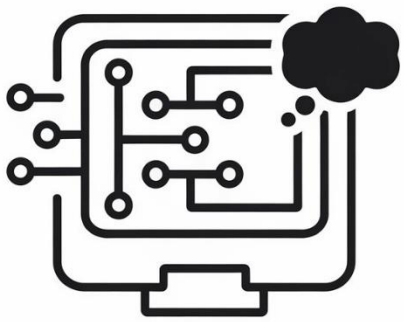

ELFMC

# Engineering Large Foundational Models for Enterprise Integration

| |
|---|
| State of the Art |

| | |
|---|---|
| Project title | Engineering Large Foundational Models for Enterprise Integration |
| Project acronym | ELFMo |
| Project number | 23004 |
| Dissemination level | Public |
| License | CC-BY 4.0 |
| Version | 1.0 |
| Date | 2025-03-06 |

**Contributors**

| | |
|---|---|
| Editor(s) | Mikko Raatikainen (University of Helsinki), Pilar Aranda (Dextromedica), Juhani Kivimäki (University of Helsinki) |
| Reviewer(s) | Mikko Raatikainen (University of Helsinki), Pilar Aranda (Dextromedica), Juhani Kivimäki (University of Helsinki) |
| Contributor(s) | Mikko Raatikainen (University of Helsinki), Niila Siilasjoki (University of Helsinki), Juhani Kivimäki (University of Helsinki), Jorma Valjakka (University of Helsinki), Juuso Saavalainen (University of Helsinki), Tuuli Lindroos (F-Secure), Fabio Román (Dextromedica), Pilar Aranda (Dextromedica) |

**Abstract**

The integration of Large Foundation Models (LFMs) and Generative AI into business, while expansive, introduces a wide array of risks and challenges due to costs, compliance issues, and technical complexities. The ELFMo project aims to address these challenges by providing a framework for effective integration, also enabling enterprises to navigate legal, security, and ethical concerns while aligning with European regulations. ELFMo empowers organizations to reliably integrate LFMs and Generative AI into their infrastructures and offerings, allowing them to maintain control over risks, challenges, and opportunities.

This report presents the state of the art regarding the targetter innovations of the ELFMo projet: Innovation 1: A risk-based approach to informed decision making for the rapid integration of LFMs into one's business environment. Innovation 2: Tools, methods and infrastructures for trustworthy adaptation and integration of LFMs to domain-specific tasks. Innovation 3: Evidence-based procedures for quality and compliance assessment for LFM-based applications and services. Innovation 4: Fostering open-source and open-access solutions and European standards.

**Table of contents**

Introduction ..... 6

Intended audience ..... 7

Related documents ..... 7

Innovation 1: A risk-based approach to informed decision making for the rapid integration of LFMs into one's business environment. ..... 8

Absence of Risk-Aware Model Selection ..... 8

Fragmented Validation and Verification Practices ..... 9

Regulatory Pressure and Compliance Challenges ..... 9

Innovation 2: Tools, methods and infrastructures for trustworthy adaptation and integration of LFMs to domain-specific tasks. ..... 10

Adaptation approaches ..... 10

Technologies ..... 12

Data Preparation and Augmentation ..... 12

Retrieval-Augmented Architectures and Vector Databases ..... 13

Workflow Orchestration and Agent Frameworks ..... 13

Observability, Experiment Tracking, and MLOps ..... 13

Adaptation and Training Methods ..... 14

Infrastructure and Operations ..... 14

Evaluation Metrics for Domain-Specific LFM Adaptation ..... 15

Implications and Limitations of the Current State of the Art ..... 17

Innovation 3: Evidence-based procedures for quality and compliance assessment for LFM-based applications and services. ..... 18

Model-Level Evaluation and Benchmarking ..... 18

Lifecycle Governance and MLOps ..... 18

Risk-Based Assessment and Regulatory Context ..... 19

Continuous Monitoring and Drift Detection ..... 19

Auditing Frameworks and Cross-Stakeholder Evidence ..... 19

Gaps in the State of the Art ..... 20

Innovation 4: Fostering open-source and open-access solutions and European standards ..... 21

Open-Source LFMs and Open-Access Resources ..... 21
Open-Source vs Open-Weight Models ..... 21
Leaderboards and Practical Evaluation Context ..... 22
Practical VRAM Tier ..... 22
Open Datasets and other resources ..... 23
Hallucination Evaluation Datasets ..... 23
Open-Source Tooling and Infrastructure ..... 24
Standards and Regulation ..... 25
References ..... 27

# Introduction

The integration of Large Foundation Models (LFMs) and Generative AI into business, while expansive, introduces a wide array of risks and challenges due to costs, compliance issues, and technical complexities. The ELFMo project aims to address these challenges by providing a framework for effective integration, also enabling enterprises to navigate legal, security, and ethical concerns while aligning with European regulations. ELFMo empowers organizations to reliably integrate LFMs and Generative AI into their infrastructures and offerings, allowing them to maintain control over risks, challenges, and opportunities.

The integration of AI is increasingly recognized as a major challenge for businesses. Despite the promises, a survey (MIT 2025) reveals that 95% of organizations are still seeing no meaningful return from their GenAI and large foundation model (LFM) investments beyond proofs of concept or pilots, largely because implementations break down due to brittle workflows, limited contextual learning, and weak alignment with day-to-day operations. The core barrier to scaling AI is therefore not infrastructure, regulation, or talent per se, but the inability to embed AI seamlessly into real business processes. Although general-purpose tools such as Microsoft Copilot and ChatGPT clearly enhance individual productivity, the broader enterprise-level impact of LFM-based solutions remains uncertain, underscoring the need for strategic, integrated adoption as both technologies and organizational capabilities mature. In fact, another survey (McKinsey 2025) indicates that the highest-performing organizations stand out by looking beyond incremental efficiency gains. They view AI as a catalyst for full organizational transformation, redesigning workflows, accelerating innovation, and fundamentally reshaping how value is created.

This report provides an overview of the current state of the art in targeted technical innovations of the ELFMo project. The report establishes a baseline from which further developments can be assessed. This baseline will serve as a critical reference point for guiding future research, supporting design decisions, and enabling the systematic development of new innovations within the ELFMo project.

The document specifically covers the state of the art for all four innovations of the ELFMo project: Innovation 1: A risk-based approach to informed decision making for the rapid integration of LFMs into one's business environment. Innovation 2: Tools, methods and infrastructures for trustworthy adaptation and integration of LFMs to domain-specific tasks. Innovation 3: Evidence-based procedures for quality and compliance assessment for LFM-based applications and services. Innovation 4: Fostering open-source and open-access solutions and European standards.

## Intended audience

The primary audience for this document is the ELFMo consortium and key project stakeholders, who will use the state-of-the-art to inform development, exploitation, and innovation within ELFMo. By outlining the state-of-the-art based baseline, this document supports informed decision-making and identifies opportunities for advancing trustworthy, European-aligned AI integration. While aimed at technically oriented participants in ELFMo's research and adoption activities, the document is public and may also benefit external readers interested in recent advances in open-source and open-access solutions for LFM integration.

## Related documents

The technical deliverables of the ELFMo project report the advances compared to this state of the art.

# Innovation 1: A risk-based approach to informed decision making for the rapid integration of LFMs into one's business environment.

Innovation 1 addresses the need for companies to be able to systematically assess the risks and opportunities associated with integrating LFM into their processes. It introduces a risk-based approach to informed decision making, where the entry point is the business use case, not the model. Each use case is associated with a multi-dimensional risk assessment (ethical, legal, operational, and domain-specific). Model selection, adaptation strategy, validation effort, and deployment architecture are derived from this risk profile. This approach aligns closely with recent academic proposals for trustworthy AI governance (Fan et al., 2023; Lipizzi, 2024) and provides a practical pathway for enterprises to reduce deployment risk, shorten decision cycles, and  and align with regulatory expectations. Innovation 1 forms the foundational layer of the ELFMo innovation stack, enabling and guiding subsequent technical innovations.

LFMs have reached a level of maturity that enables rapid experimentation and deployment across a wide range of enterprise use cases. Through cloud-based APIs and open-source frameworks, organizations can quickly prototype generative AI applications for task such as document processing, customer interaction, decision support, and knowledge management. However, this rapid accessibility masks a lack of structured decision-making frameworks, especially in domains where errors, bias, or lack of transparency may have severe consequences.

Current state-of-the-art enterprise adoption practices primarily rely on technology-centric criteria, such as benchmark performance, inference cost, latency, and ecosystem availability (Liang et al., 2022). While these criteria are useful for early-stage exploration, they are insufficient for informed decision making in regulated and high-risk environments such as healthcare

## Absence of Risk-Aware Model Selection

Existing approaches to LFM adoption rarely incorporate explicit risk modeling. Model selection is often based on static benchmark results (e.g. MMLU, HELM, BIG-bench), which do not account for the specific context of use, fail to capture ethical, legal and operational risks, and are largely disconnected from deployment realities. Furthermore, many commercial LFMs provide limited transparency regarding training data provenance, update policies, and internal alignment mechanisms. This opacity poses a major challenge in sensitive sectors such

as healthcare, where traceability, reproducibility, and accountability are mandatory (Sun et al., 2024).

## Fragmented Validation and Verification Practices

Although research proposes comprehensive taxonomies of risks associated with LLM systems (Cui et al., 2024), these frameworks remain largely theoretical and are rarely translated into operational decision-support tools for enterprises. Validation and verification (V&V) practices typically focus on isolated model performance, offline benchmarks, or limited human evaluation. Instead, they seldom address system-level behavior, continuous model evolution, or deployment-specific constraints. As highlighted in the medical use case in the EFLMo project, this gap becomes critical when LFMs are used to support administrative, or operational decisions in healthcare or similar high-risk environments.

## Regulatory Pressure and Compliance Challenges

The introduction of the EU Artificial Intelligence Act reinforces the need for a shift toward risk-based AI governance. Under the AI Act, some LFM-based applications, particularly those related to healthcare decision support, are expected to fall under *high-risk* classifications (European Commission, 2024). Compliance requires, e.g., documented risk assessments, demonstrable mitigation strategies, and traceability across the AI lifecycle. However, the currrent state of the art lacks integrated methodologies that allow organizations to balance rapid innovation with regulatory compliance, particularly for SMEs with limited resources.

# Innovation 2: Tools, methods and infrastructures for trustworthy adaptation and integration of LFMs to domain-specific tasks.

This innovation addresses the need for enterprises to translate risk-aware decisions about Large Foundation Models (LFMs) into concrete, deployable, and trustworthy technical solutions. While current state-of-the-art approaches provide a wide range of adaptation techniques and infrastructure components, they are typically fragmented and lack an integrated perspective on trustworthiness, lifecycle management, and regulatory compliance. Within the ELFMo project, this innovation fills this gap by introducing a systematic toolbox of tools, methods, and infrastructures designed to support the trustworthy adaptation and integration of LFMs into domain-specific tasks, particularly in regulated and high-impact environments such as healthcare and telecommunications.

This approach is closely aligned with recent research on trustworthy and responsible AI systems, which emphasizes the integration of governance, validation, and monitoring mechanisms throughout the AI lifecycle rather than as post-deployment controls (Fan et al., 2023; Cui et al., 2024; Sun et al., 2024). It also directly supports the requirements of emerging regulatory frameworks, such as the EU Artificial Intelligence Act, which mandate systematic risk mitigation, transparency, and lifecycle traceability for high-risk AI systems (European Commission, 2024).

By operationalizing the risk-based decisions defined in Innovation 1, Innovation 2 provides enterprises with practical means to safely adapt general-purpose LFMs to domain-specific tasks, deploy and operate them across heterogeneous infrastructures, reduce integration and operational risks, and accelerate the transition from experimentation to production while maintaining regulatory alignment. Together, Innovations 1 and 2 constitute the core methodological and technical foundation of the ELFMo innovation stack, enabling subsequent advances in benchmarking, validation, and continuous learning.

## Adaptation approaches

The recent evolution of Large Foundation Models (LFMs), and particularly Large Language Models (LLMs), has marked a paradigm shift in artificial intelligence, enabling models to perform a wide variety of cognitive tasks with minimal task-specific training. Pre-trained on massive and heterogeneous corpora, these models exhibit strong generalization capabilities across domains such as text understanding, reasoning, summarization, and dialogue (Brown et al., 2020; Bommasani et al., 2021).

Domains such as healthcare, telecommunications, customer service, and industrial decision support, explicitly addressed within the ELFMo project pose stringent requirements in terms of reliability, traceability, data governance, explainability, and regulatory compliance. The main challenge is not merely improving model accuracy, but ensuring that adapted LFMs behave in a trustworthy, auditable, and controllable manner throughout their operational lifecycle.

Current industrial adoption of LFMs often relies on rapid prototyping using commercial APIs or open-source models, guided primarily by benchmark performance, latency, and cost considerations (Liang et al., 2022). However, such criteria provide only a partial view of model suitability and fail to capture domain-specific risks and operational constraints, especially in regulated sectors. In healthcare, for example, incorrect or biased outputs may have direct clinical consequences, while insufficient traceability or transparency can lead to non-compliance with legal and ethical standards (Sun et al., 2024).

The limitations in the state of the art include the following.

- Fragmentation of adaptation techniques: Prompt engineering, RAG, and fine-tuning are typically applied in isolation, without a unified framework guiding their selection, combination, and validation according to domain risk.
- Insufficient integration of trust mechanisms: Security, bias mitigation, validation, and explainability are often external layers rather than intrinsic components of adaptation pipelines.
- Infrastructure heterogeneity and operational complexity: Enterprises must deploy LFMs across diverse environments (cloud, on-premise, edge), yet SotA solutions rarely provide coherent orchestration, monitoring, and routing mechanisms.
- Lack of lifecycle traceability: Continuous updates, fine-tuning iterations, and inference-time decisions are poorly documented, complicating auditing and regulatory compliance, particularly under frameworks such as the EU AI Act.

These limitations are especially critical in regulated environments, such as healthcare, where the medical use cases demonstrate that model adaptation choices directly impact safety, ethical compliance, and legal responsibility.

The ELFMo project addresses these challenges through its second innovation pillar, dedicated to providing a technical and methodological toolbox for the trustworthy adaptation and integration of LFMs into domain-specific tasks. This toolbox combines software tools, adaptation methods, and scalable infrastructures designed to support both experimental validation and enterprise-grade deployment, while embedding trustworthiness by design.

# Technologies

ELFMo leverages a combination of open-source technologies and partner-driven developments to enable flexible, reproducible, and controlled adaptation workflows. The adaptation and integration of Large Foundation Models (LFMs) into domain-specific applications is supported by a rapidly evolving ecosystem of open-source software tools and libraries. These tools span model access and training, data preparation, retrieval mechanisms, workflow orchestration, and lifecycle management. While heterogeneous and often loosely integrated, they collectively constitute the current state of the art for operationalizing LFMs in enterprise environments.

At the core of the LFM ecosystem lies the Hugging Face Transformers library, which has become a de facto standard interface for accessing and adapting pre-trained transformer-based models (Wolf et al., 2020). The library provides standardized APIs for loading and fine-tuning models, support for distributed training, compatibility with PyTorch and TensorFlow, and integration with model hubs hosting thousands of open-weight models.

In recent years, parameter-efficient fine-tuning (PEFT) approaches have gained prominence as a scalable alternative to full-model retraining. Among these, LoRA (Low-Rank Adaptation) introduces trainable rank-decomposition matrices into attention layers, enabling adaptation while modifying only a small fraction of model parameters (Hu et al., 2021). Its quantized extension, QLoRA, further reduces memory requirements by combining low-rank adaptation with 4-bit quantization, allowing large models to be fine-tuned on consumer-grade hardware (Dettmers et al., 2023). These approaches significantly lower computational costs while preserving performance, making them particularly relevant for enterprise adoption.

In addition to fine-tuning methods, knowledge distillation techniques are widely employed to transfer capabilities from large LFMs to smaller, more efficient models (Hinton et al., 2015). Distillation enables the development of Small Language Models suitable for latency-sensitive or resource-constrained environments, such as edge deployment scenarios.

## Data Preparation and Augmentation

Data quality remains a critical determinant of successful domain adaptation. The state of the art includes multiple toolkits designed to automate preprocessing tasks such as text normalization and cleaning, deduplication, anonymization and privacy filtering, and semantic enrichment and tagging.

Synthetic data generation techniques are increasingly applied to mitigate data scarcity and privacy limitations. These techniques include rule-based augmentation, back-translation, paraphrasing via LLMs, and controlled data synthesis, enabling dataset expansion without exposing sensitive information.

Recent research highlights the risks associated with recursively training on synthetic data (Shumailov et al., 2023), emphasizing the importance of carefully governed augmentation strategies. Consequently, modern data pipelines increasingly incorporate validation and filtering stages to ensure consistency and prevent degradation of model performance.

## Retrieval-Augmented Architectures and Vector Databases

Retrieval-Augmented Generation (RAG) has emerged as one of the dominant architectural paradigms for grounding LFMs in domain-specific knowledge (Lewis et al., 2020). RAG systems combine parametric knowledge encoded in the model with non-parametric knowledge retrieved from external sources at inference time.

Key components of RAG architecture include embedding models for semantic representation, vector databases (e.g., Chroma, Pinecone, Milvus), and similarity search mechanisms for document retrieval. By incorporating dynamically retrieved context, RAG architectures reduce hallucinations and improve factual consistency, particularly in knowledge-intensive domains.

The growing availability of production-ready vector databases has facilitated enterprise adoption of RAG pipelines. These databases provide scalable indexing, low-latency retrieval, and integration with modern orchestration frameworks.

## Workflow Orchestration and Agent Frameworks

To operationalize LFMs within complex enterprise workflows, orchestration layers have become essential. Frameworks such as LangChain and LlamaIndex enable modular construction of pipelines involving prompt templates, retrieval modules, tool usage and API calls, multi-step reasoning chains, and agent-based decision logic.

Such frameworks support composability and rapid experimentation, allowing developers to integrate LFMs into multi-component systems. However, they also introduce new challenges related to traceability, debugging, and security, particularly when combined with external tools and dynamic memory mechanisms.

## Observability, Experiment Tracking, and MLOps

The operationalization of LFMs increasingly relies on MLOps and LLMOps platforms to manage experimentation and deployment. Widely adopted tools include:

- MLflow (Zaharia et al., 2018) for experiment tracking and model versioning,
- Kubeflow for orchestrating machine learning pipelines,
- Weights & Biases for hyperparameter logging and performance monitoring.

More recent observability platforms focus specifically on LLM-based systems, tracking prompt inputs, outputs, latency, token usage, and error patterns. These tools aim to provide transparency into model behavior in production settings, supporting reproducibility and regulatory compliance.

Despite these advances, the state of the art remains fragmented. Tools are often loosely coupled, requiring substantial engineering effort to integrate them into coherent, trustworthy pipelines suitable for regulated environments.

## Adaptation and Training Methods

To ensure effective performance in domain-specific contexts, ELFMo combines complementary adaptation strategies.

Techniques such as LoRA (Low-Rank Adaptation) and QLoRA are prioritized, enabling re-training of less than 1% of model parameters while reducing computational requirements by approximately 90–95%. This approach supports frequent updates and experimentation without the cost and risk of full model retraining.

The ELFMo project applies *multi-shot prompting*, *Chain-of-Thought* reasoning, and automated optimization techniques such as Automatic Prompt Engineer (APE) to guide model behavior without modifying base model weights. External domain knowledge is integrated through vector databases (e.g., ChromaDB, Pinecone), grounding model outputs in curated and up-to-date information sources. This significantly reduces hallucinations and improves factual accuracy in high-stakes domains. ELFMo explores the use of Small Language Models, trained via knowledge distillation from larger LFMs, for high-volume and narrowly defined tasks where latency, cost efficiency, and energy consumption are critical.

## Infrastructure and Operations

The ELFMo infrastructure supports the complete lifecycle of LFM adaptation and deployment such as vLLM optimized for GPU clusters with tensor parallelism and high-throughput inference and llama.cpp, a lightweight C++ backend enabling inference on CPU-only environments and edge devices.

Kubernetes is used for elastic scaling and service orchestration, while SLURM supports efficient job scheduling across heterogeneous hardware infrastructures.

A dedicated inference router dynamically distributes requests across multiple models (“experts”) based on criteria such as domain affinity, complexity, or performance constraints (e.g., routing queries to domain-specialized models).

Integration with MLflow, Kubeflow, and Weights & Biases enables systematic logging of metrics, model versions, configurations, and experimental results, ensuring reproducibility, auditability, and regulatory readiness.

## Evaluation Metrics for Domain-Specific LFM Adaptation

The evaluation of adapted Large Foundation Models (LFMs) in enterprise and regulated environments has evolved beyond traditional NLP accuracy metrics. While early large language model research primarily focused on benchmark performance (e.g., MMLU, BIG-bench, GLUE), the operational deployment of LFMs in domain-specific contexts requires multi-dimensional evaluation frameworks that incorporate reliability, factual consistency, efficiency, cost, and trustworthiness (Liang et al., 2022; Bommasani et al., 2021).

In high-impact domains such as healthcare and telecommunications, evaluation must reflect not only task-level correctness but also operational and regulatory constraints.

At the core of LFM evaluation traditional task-specific metrics remain, including, accuracy and F1-score for classification tasks, exact match for question answering, BLEU and ROUGE for text generation and semantic similarity metrics based on embeddings

These metrics provide quantitative measures of task performance but are insufficient to capture risks such as hallucination, bias, or instability in real-world deployment scenarios.

**Hallucination and Factual Consistency Metrics**

Hallucinations—defined as plausible but factually incorrect outputs—represent one of the primary risks in domain-specific LFM deployment. The state of the art includes both human and automated approaches to hallucination assessment.

Automated methods include:

- FActScore, which decomposes generated text into atomic facts for factual verification (Min et al., 2023)
- TruthfulQA, designed to evaluate model resistance to common misconceptions (Lin et al., 2021)
- Natural Language Inference (NLI)-based entailment checks to assess factual alignment

In operational systems, hallucination rates are often aggregated to produce average hallucination indicators across evaluation runs, reflecting model reliability in knowledge-intensive contexts.

**Latency and Operational Efficiency Metrics**

Enterprise deployment requires evaluation of run-time performance indicators, including average response latency, throughput (requests per second), token generation speed and batch processing efficiency.

LLMOps research emphasizes the importance of latency monitoring and inference profiling for production systems (Zhang et al., 2023). In time-sensitive environments, such as clinical support or customer interaction systems, response delay may significantly impact usability and safety.

**Cost and Resource Utilization Metrics**

With the rise of API-based and GPU-intensive LFM deployment, cost has become a first-class evaluation dimension. Current practice includes measuring cost per inference run, cost per 1,000 tokens, GPU memory utilization and energy consumption estimates.

Parameter-efficient methods such as LoRA and QLoRA have demonstrated substantial reductions in computational requirements (Hu et al., 2021; Dettmers et al., 2023), but cost-performance trade-offs remain context-dependent. Evaluating economic efficiency alongside technical performance is therefore increasingly recognized as essential for enterprise adoption.

**Multi-Dimensional Trustworthiness Indicators**

Research proposes composite evaluation frameworks for assessing broader trustworthiness dimensions of LFMs (Sun et al., 2024; Cui et al., 2024). These typically include robustness to adversarial inputs, bias and fairness indicators, safety and harmful output detection, consistency across repeated queries, and explainability and interpretability proxies.

Visualization techniques such as radar or radial plots are commonly used to represent trade-offs between these dimensions, enabling comparative assessment across models.

**Emerging Use of LLM-as-a-Judge**

Another evaluation paradigm involves the use of one LLM to assess the outputs of another, commonly referred to as LLM-as-a-judge (Zheng et al., 2023). This approach is increasingly applied to evaluate response relevance, coherence, instruction following and comparative quality between models.

While promising in scalability, this method introduces additional layers of uncertainty and bias, requiring careful calibration and validation.

## Implications and Limitations of the Current State of the Art

The current ecosystem of tools, methods, and infrastructures for LFM adaptation has significantly reduced the technical barriers to experimentation and deployment. Parameter-efficient fine-tuning methods such as LoRA and QLoRA have lowered computational requirements (Hu et al., 2021; Dettmers et al., 2023), retrieval-augmented architectures have improved factual grounding (Lewis et al., 2020), and scalable orchestration platforms have facilitated distributed inference and model lifecycle management.

However, despite these advances, several structural limitations remain:

- Resource optimization remains context-dependent: While PEFT and quantization techniques reduce memory and computational requirements, their efficiency gains vary significantly depending on hardware configuration, model size, and domain complexity.
- Time-to-deployment is highly engineering-dependent: Although modern frameworks accelerate prototyping, transitioning from experimentation to robust production systems still requires substantial integration effort, particularly in regulated environments.
- Digital sovereignty remains an open challenge: The increasing availability of open-weight models has improved accessibility, yet many high-performance models remain controlled by non-European providers, raising concerns regarding dependency, governance, and compliance with European regulatory frameworks (Bommasani et al., 2021).

As a result, while the state of the art provides powerful building blocks, it does not yet offer a fully integrated, risk-aware, and regulation-ready framework for trustworthy domain-specific LFM deployment.

# Innovation 3: Evidence-based procedures for quality and compliance assessment for LFM-based applications and services.

Assurance of quality, safety, and regulatory compliance are core components of trustworthy AI. With emerging legislation, such as the EU AI Act, which imposes conformity assessments, risk mitigation, and ongoing monitoring requirements for high-risk and *systemic* AI models, foundation models (LFMs) are increasingly subject to formal compliance obligations. LFMs such as large language models cannot simply be treated as traditional software or small ML models due to their complexity, scale, and deployment patterns — leading to novel testing, auditing, and monitoring challenges not yet satisfactorily addressed by existing standards or tools. (Reuters, 2025)

## Model-Level Evaluation and Benchmarking

The state of practice for assessing LFMs has evolved rapidly over the past few years. Research on LFM evaluation has typically focused on *benchmarking and safety testing*, where the model's capacity across tasks and potential harmful tendencies (bias, toxicity, hallucination) have been measured using curated datasets and adversarial testing protocols (Bommasani, 2021). This research highlights the need for robust pre-deployment evaluation procedures that extend beyond narrow task performance (NIST, 2023).

Documentation formats such as model cards (Mitchell, 2019) have been proposed to summarize model intent, limitations, and metrics at a high level. While these artifacts improve transparency, they function primarily as *static disclosures* and do not constitute mechanisms for continuous, real-time evidence generation. They also do not inherently link model behavior to *application-specific risk profiles*.

## Lifecycle Governance and MLOps

Industry and enterprise AI governance frameworks extend traditional software lifecycle practices to machine learning operations (MLOps), incorporating automated testing, continuous integration/continuous deployment (CI/CD), and drift monitoring. Broad-scope standards (ISO, 2023) specify requirements for *AI management systems* covering governance, risk assessment, documentation, and monitoring across the AI lifecycle.

Similarly, the NIST AI Risk Management Framework (AI RMF) provides a structured approach to managing risk through four iterative core functions — *Govern*, *Map*, *Measure*, and *Manage* (NIST, 2023). The Generative AI profile within the AI RMF provides further guidance to foundation models, emphasizing traceability, misuse mitigation, and continuous evaluation. Even so, these lifecycle frameworks are *principle-based and high level*, often insufficient to operationalize domain-specific monitoring or evidence that aligns with regulatory compliance

requirements. Furthermore, successful governance requires finding a balance between automated and human-in-the-loop approaches to ensure trustworthiness (Tanna, 2025).

## Risk-Based Assessment and Regulatory Context

Risk-based governance is widely advocated for AI assurance. Organizations such as the OECD promote risk management principles that integrate ethical, legal, and safety dimensions across an AI system's life (OECD, 2024). Analyses show that traditional standards and checklists must evolve to include *model and data quality, process validation, and decision transparency* and contain dynamic aspects such as continuous monitoring and explainability (Szadeczky, 2025).

The EU AI Act, now legally in force, explicitly requires technical documentation, evaluation of systemic risks, incident reporting, and periodic re-evaluation of high-risk AI systems (including foundation models). This creates an imperative for evidence that can demonstrate compliance with specific Articles (e.g., risk assessment, post-market monitoring, and technical documentation obligations), but the act itself does not prescribe detailed audit procedures, leaving implementation to stakeholders (Reuters, 2025)

## Continuous Monitoring and Drift Detection

Traditional machine learning monitoring focuses on tracking model performance and data integrity over time, using metrics such as accuracy, error rates, or statistical comparisons between production and training data to detect issues such as data or concept drift. Such methods assume well-defined input–output relationships, narrow task definitions, and full access to training distributions, and are typically embedded in MLOps pipelines with automated adaptation triggers. These approaches have been widely surveyed and form the basis of current enterprise model monitoring practices (Schröder, 2022).

However, LFMs introduce new failure modes and monitoring requirements. Their outputs are generative and semantically rich, task and domain adaptation is widespread, and base models are often opaque or hosted externally, limiting access to training data and model internals. Instead of relatively easily quantifiable drifts in performance, model issues manifest as problems in factuality, desired style, or safety (Huang, 2025). Also, accountability is distributed across model providers, integrators, and application owners. Existing monitoring toolkits address only fragments of these challenges — for example performance tracking, drift detection, or semantic evaluation — and no current framework provides a comprehensive closed-loop solution that integrates continuous monitoring with actionable remedies on the ML system level.

## Auditing Frameworks and Cross-Stakeholder Evidence

Audit practices for AI systems vary across industries and governance bodies. Recent AI auditing frameworks and guidelines (e.g. IIA, 2024; ECIIA, 2052) emphasize governance controls and continuous oversight, mirroring many risk and compliance control principles,

especially regarding General Purpose AI (GPAI) models, which are LFMs by definition. Independent academic work also explores *continuous auditing* where frameworks assess a system's performance against a set of criteria over time rather than at a single point — an approach better aligned with regulation-driven compliance audition (Minkkinen, 2022).

Multi-stakeholder approaches are needed to reconcile the tension between audit transparency and proprietary model confidentiality (Mökander, 2022). Emerging cryptographic approaches, such as *zero-knowledge auditing*, attempt to demonstrate new directions for verifiable compliance without exposing sensitive intellectual property (Scaramuzza, 2025).

## Gaps in the State of the Art

Despite progress in benchmarking, governance frameworks, and lifecycle risk management, several gaps remain:

- Lack of integrated solutions that unify model-level audit evidence with application-specific risk profiles and compliance reporting.
- Absence of widely adopted continuous monitoring standards tailored specifically to LFM behavior in enterprise contexts with a clear link to business KPIs.
- Insufficient tooling that links automated technical metrics (drift, hallucination) with compliance criteria required under regulatory frameworks such as the EU AI Act.

The proposed evidence-based procedures directly address these gaps by establishing a *continuous, risk-based auditing and monitoring framework* that spans the model lifecycle and operational contexts of LFM applications. By enabling coordinated evidence generation — from model provider disclosures to application-level risk assessment — and structuring this evidence for regulatory conformity demonstration, this innovation goes beyond current benchmarking and governance practices, extending the frontier of enterprise LFM quality and compliance assurance.

# Innovation 4: Fostering open-source and open-access solutions and European standards.

The focus of ELFMo Innovation 4 is fostering open-source and open-access solutions and European and industrial standards. While the ELFMo project is not aiming at developing own LLM/LFM models, we summarize in this section numerous open solutions and elaborate the most compelling concrete examples that can utilized in LFM-based systems as well as contributions can be made for improvements.

## Open-Source LFMs and Open-Access Resources

LFMs were popularized by proprietary models accessible through web interfaces and APIs, the best-known example being probably OpenAI's ChatGPT. Numerous open-source alternatives have since emerged. The open models originate from different players such as research labs, open-source foundations, academic institutions, corporate contributors, and community-driven groups.

The most compelling examples are currently found from model families such as Llama 4, gpt-oss, Qwen 3, Mistral 3, and DeepSeek (V3 & R1). These model families represent the current state of the art in open and open-weight models. It's worth noting that the exact models vary depending on the category, such as reasoning, logic, coding, multilingual, math, tooling, agentic capabilities, efficiency, and metrics used to define the performance. Models in general move fast and updated versions of existing model families and competing models come out in rapid phase.

## Open-Source vs Open-Weight Models

Open-weight models publish trained parameters (weights) so others can run inference and often fine-tune locally, but they may not release full training data, full training code, or complete provenance. (Open Source Initiative, n.d. Open-source (OSI sense) requires compliance with the Open Source Definition. (Open Source Initiative, n.d.)

Examples of open-weight, not OSI-open-source, offerings currently are: Meta's Llama 3.3 70B and Google's Gemma 2 (9B) both release weights under custom, or restricted licenses rather than an OSI-approved license. Llama 3.3 requires agreement to Meta's community license terms (including an Acceptable Use Policy) rather than a permissive OSI license. (Meta via Hugging Face, 2024) Gemma 2 similarly requires accepting Google's usage license/terms. (Google via Hugging Face, n.d.; Google, n.d.)

In contrast, Mistral's models (including Mixtral) represent prominent European open-weight systems: while their weights are publicly released and widely used, they are distributed under licenses that do not meet OSI open-source requirements and therefore fall outside the OSI definition of open source (Mistral AI, 2025). Another notable European open model is Poro 34B, released under Apache 2.0 and trained with LUMI/EuroHPC resources, targeting Finnish and English languages (LumiOpen, 2024).

## Leaderboards and Practical Evaluation Context

Public leaderboards are commonly used to compare the performance of large language models under standardized evaluation settings. While they provide a convenient snapshot of relative performance, they should be interpreted with caution, as results may depend on prompt design, evaluation protocols, and benchmark selection.

The Hugging Face Open LLM Leaderboard, which is currently the leading service, evaluates models across a fixed set of benchmarks, focusing primarily on automated, reproducible metrics such as reasoning, knowledge, and coding performance (Hugging Face, n.d.).

The LMSYS Chatbot Arena (LMArena, 2025) adopts a different approach, relying on crowd-sourced human preference comparisons in which users evaluate pairs of anonymous model outputs. This setting captures subjective qualities such as helpfulness and conversational quality but is sensitive to prompt distribution and sampling effects (LMArena, n.d.).

## Practical VRAM Tier

Hardware constraints strongly influence which LFMs are practically usable in research and deployment. In particular, available GPU memory limits model size, context length, and achievable precision.

- 8–12 GB VRAM: Models around 7B–9B can often run with 4-bit quantization, with practical limits determined by KV-cache/context and serving overhead. Examples include Qwen2.5-7B-Instruct and Gemma 2 9B. (Qwen via Hugging Face, n.d.; Google via Hugging Face, n.d.)

- 16–32GB VRAM: Models like Phi-4 (14B) and Mistral Small 3 (24B) become practical with 4-bit or 8-bit quantization, again depending on context/KV cache and runtime overhead. (Microsoft via Hugging Face, 2024; Mistral AI, 2025)
- 80 GB+ VRAM (A100/H100 class): Sufficient to load 70B-class models comfortably at higher precision and/or with larger context windows, subject to runtime overhead.

## Open Datasets and other resources

Benchmark datasets play a central role in evaluating LLM capabilities, particularly for reasoning, mathematics, coding, and factual question answering. Commonly cited and used benchmark datasets today include:

- Massive Multitask Language Understanding (MMLU) is a benchmark dataset for evaluating the performance of Large Language and other NLP models. It contains multiple-choice questions and answers from 57 domains. It has also inspired more challenging variations, such as MMLU-Pro, which include more challenging questions. (Hendrycks et al., 2020; Wang et al., 2024)
- Grade School Math 8k (GSM8k) consists of verbal grade school math questions requiring multi-step reasoning. (Cobbe et al., 2021)
- MATH is a set of 12.5 thousand high school competition-level challenging mathematics problems. (Hendrycks et al., 2021)
- Ai2-ARC is a dataset containing grade school multiple-choice science questions. The questions are segmented into easy and challenging questions based on the ability of retrieval-based and word co-occurance algorithms' ability to answer them. (Clark et al., 2018)
- GPAQ is a set of challenging expert-level multiple-choice questions in domains of biology, physics and chemistry. It has been designed to test performance of models specifically on tasks requiring deep level of topic expertise as well as reasoning abilities. (Rein et al., 2024)
- Mostly Basic Programming Problems (MBPP) is a dataset consisting of a thousand entry-level programming questions, solutions in Python and tests. (Austin et al., 2021)
- MT Bench is a set of multi-turn questions specifically tailored for evaluating LLM-based chat assistants. (Zheng et al., 2023)
- PubMedQA is a dataset with biomedical questions and answers sourced from PubMed abstracts. A small part of the question-answer pairs is expert-annotated, but majority is artificially generated. For each question, the dataset contains context, verbose answer and a yes/no/maybe conclusion. (Jin et al., 2019)

## Hallucination Evaluation Datasets

Hallucination-specific benchmarks are designed to probe failure modes where models generate fluent but factually incorrect outputs. Commonly Cited Hallucination Evaluating Datasets include:

- TruthfulQA is a question-answering dataset designed to probe cases where models confidently produce false but plausible-sounding answers rooted in common misconceptions. (Lin et al., 2022)
- HaluEval is a large-scale hallucination evaluation benchmark with human annotation. (Li et al., 2023)
- TriviaQA is a large dataset with question–answer pairs. (Joshi et al., 2017)
- HotpotQA is a multi-hop QA dataset where questions require reasoning over multiple supporting documents, with sentence-level supporting facts. (Yang et al., 2018)
- BEIR (Thakur et al., 2021) is a large-scale heterogeneous benchmark for zero-shot evaluation of information retrieval models introduced by Thakur et al. in 2021. It is designed to test how well retrieval systems generalize out of domain rather than only in settings like their training data. It evaluates information retrieval systems across a wide range of domains and query types, making it well suited for assessing retrieval robustness in sparse and fact-oriented scenarios.
- LoTTE is a benchmark for long-tail, topic-stratified evaluation of retrieval systems, first introduced within the ColBERTv2architecture paper by Santhanam et al. (2022). LoTTE is a benchmark dataset focused on long-tail, domain-specific queries, which pose challenges for contextual relevancy and grounding in retrieval-augmented generation systems.

## Open-Source Tooling and Infrastructure

Open-source evaluation frameworks such as DeepEval, RAGAS, and HELM play a central role in assessing the quality and reliability of large language model (LLM) and retrieval-augmented generation (RAG) systems. DeepEval has gained particular relevance due to its focus on fine-grained, metric-driven assessment of retrieval-augmented generation (RAG) pipelines. DeepEval provides standardized evaluators for answer relevancy, faithfulness, contextual precision, and hallucination detection, and supports LLM-based judging to assess semantic alignment between retrieved context and generated outputs. These capabilities make it well suited for diagnosing failure modes within RAG pipelines at both the retrieval and generation stages. RAGAS emphasizes the relationship between retrieved context and generated responses through grounding and relevance metrics. HELM offers a broader benchmarking perspective, evaluating models across multiple dimensions including robustness and efficiency on standardized datasets. Although these tools enable systematic and reproducible evaluation, they are primarily designed for offline or post-hoc analysis and assume static pipeline configurations.

Deploying LLMs at production involves multiple architectural layers. The hardware and Kernels are the foundational compute layer; this includes hardware drivers and custom

kernels. Low-level hardware acceleration is performed at this layer. The core inference layer where the model is executed, managing computation and memory resources. vLLM, DeepSpeed Inference, TensorRT-LLM, SGLang, and llama.cpp fall in this layer. Each of the frameworks optimize for different aspects of the deployment, such as paged attention computations (vLLM), model parallelism (DeepSpeed), reducing latency and throughput for complex agentic workflows (SGLang), accessible hardware platforms e.g CPU based platforms (llama.cpp). Inference serving engines provide a serving runtime, handling batching, and providing multi-framework support, such solutions include NVIDIA Triton, and NVIDIA Dynamo. This layer of the architecture provides GPU resource management. The orchestration and API layer of the architecture manages deployment, scaling of the solution, routing to provide load balancing at the service layer. This layer may also include the user interface. A key technology in this layer is KServe or other Kubernetes-native deployment and orchestration frameworks.

Ray is a more general-purpose computing framework for Python based ML workloads. It enables interoperable distributed computing, where developers can move the execution of their Python code from personal computers to cloud platforms and supercomputers. This makes Ray with the support of other open-source tools reduce development requirements of complex ML workflows and products.

Agent technologies are a recent orchestration layer that turns LFMs into multi-step workflows. In these workflows agents use external tools, such as retrieval, databases, APIs, and delegate subtasks to other agents, while maintaining context, and supporting human-in-the-loop approvals and auditing. Open-source building blocks include frameworks such as LangGraph, Langflow and Microsoft AutoGen. In addition, open standards for agent communication technologies have emerged, A2A and MCP being popular examples.

## Standards and Regulation

The European regulatory for AI and LFMs is primarily shaped by the AI Act and the General Data Protection Regulation (GDPR). The AI Act (formally adopted in 2024) establishes a risk-based framework for AI systems, while the GDPR governs the processing of personal data, including data used to train or operate AI systems. However, the EU AI regulatory landscape extends beyond these two instruments. Several additional regulations are relevant to AI-based applications. For example, the Data Governance Act facilitates trusted data sharing across the EU, which is particularly important for AI development and access to high-quality datasets. The Digital Markets Act (DMA) and the Digital Services Act (DSA) also affect AI-driven platforms by regulating large online intermediaries, algorithmic transparency, competition, and systemic risks. Furthermore, sector-specific legislation plays a significant role. The medical domain is a prime example, as it has long been governed by dedicated regulatory

frameworks such as the Medical Device Regulation (MDR) and the In Vitro Diagnostic Regulation (IVDR).

Although this regulatory architecture is largely in place, its practical application and interpretation—particularly regarding the interaction between the AI Act and existing sectoral legislation—are still evolving. Many compliance obligations will depend on forthcoming harmonised standards, guidance from the European Commission, and national supervisory practices.

Beyond formal regulation, the integration of large foundational models into enterprise offerings faces a critical sustainability paradox: AI could consume energy equivalent to entire countries like Argentina by 2027 (Holistic AI, 2024), while simultaneously being essential for automated ESG analysis and compliance reporting that regulators increasingly mandate through frameworks like ISSB and CSRD (Pulsora, 2025). The primary challenges include massive computational requirements for training and inference, lack of standardized environmental impact metrics, and the fact that 75% of organizations don't know what ESG standards to follow (Unit4, 2025). Industry responses are emerging across multiple fronts: technical innovations like MIT's Clover system that reduces carbon intensity by 90% through intelligent model selection (Raconteur, 2025), infrastructure approaches including renewable-powered data centers and federated learning to reduce centralized computing needs (Association for Computing Machinery, 2024), and governance initiatives such as the Green Software Foundation's lifecycle management framework (Green Software Foundation, 2025) and the Coalition for Environmentally Sustainable AI with over 100 partners (United Nations Environment Programme, 2025). Enterprises are increasingly adopting "sustainability-by-design" principles, implementing Sustainability Impact Assessments during model development, optimizing algorithms through techniques like pruning and quantization, and using task-specific rather than general-purpose models, while frameworks for measuring and reporting AI's environmental impact are being standardized through organizations like IEEE and CNCF to enable informed decision-making about the trade-offs between model performance and environmental cost.

# References

Association for Computing Machinery (2024) Green federated learning: A new era of green aware AI. ACM Computing Surveys. Available at: https://dl.acm.org/doi/10.1145/3718363

Austin, J. et al. (2021) 'Program synthesis with large language models', arXiv preprint arXiv:2108.07732.

Thakur, N. et al. (2021) BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv:2104.08663v4.

Bommasani, R. et al. (2021) On the opportunities and risks of foundation models. arXiv:2108.07258.

Clark, P. et al. (2018) 'Think you have solved question answering? Try ARC, the AI2 reasoning challenge', arXiv preprint arXiv:1803.05457.

Cobbe, K. et al. (2021) 'Training verifiers to solve math word problems', arXiv preprint arXiv:2110.14168.

Santhanam, K. et al. (2021) ColBERTv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint.

ECIIA (2025) The AI Act: Road to compliance. A Practical Guide for Internal Auditors. Available at: https://www.eciia.eu/2025/01/master-the-ai-act-eciias-essential-guide-for-internal-auditors/

Google via Hugging Face (n.d.) google/gemma-2-9b. Available at: https://huggingface.co/google/gemma-2-9b

Google (n.d.) Gemma Terms of Use. Available at: https://ai.google.dev/gemma/terms

Green Software Foundation (2025) Green AI position paper, 20 May. Available at: https://greensoftware.foundation/articles/green-ai-position-paper/

Hendrycks, D. et al. (2020) 'Measuring massive multitask language understanding', arXiv preprint arXiv:2009.03300.

Hendrycks, D. et al. (2021) 'Measuring mathematical problem solving with the MATH dataset', arXiv preprint arXiv:2103.03874.

Holistic AI (2024) AI and ESG: Understanding the environmental impact of AI and LLMs. Available at: https://www.holisticai.com/blog/environmental-impact-ai-llms

Huang, L. et al. (2025) 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', ACM Transactions on Information Systems, 43(2), pp. 1–55.

IIA Institute of Internal Auditors (2024) AI Auditing Framework (updated 2024). Available at: https://www.metamindz.co.uk/post/5-ai-auditing-frameworks-for-compliance

ISO (2023) ISO/IEC 42001:2023 Artificial Intelligence Management Systems – Requirements and Guidance. Available at: https://www.iso.org/standard/42001

Jin, Q. et al. (2019) 'PubMedQA: A dataset for biomedical research question answering', in EMNLP–IJCNLP 2019 Proceedings.

Joshi, M. et al. (2017) 'TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension', ACL. Available at: https://aclanthology.org/P17-1147/

Lin, S., Hilton, J. and Evans, O. (2022) 'TruthfulQA: Measuring how models mimic human falsehoods', ACL. Available at: https://aclanthology.org/2022.acl-long.229/

Li, J. et al. (2023) 'HaluEval: A large-scale hallucination evaluation benchmark for large language models', EMNLP. Available at: https://aclanthology.org/2023.emnlp-main.397/

LMArena (n.d.) Chatbot Arena Leaderboard. Available at: https://lmarena.ai/leaderboard

LumiOpen (2024) Poro-34B model card. Available at: https://huggingface.co/LumiOpen/Poro-34B

Hugging Face (n.d.) Open LLM Leaderboard: About. Available at: https://huggingface.co/docs/leaderboards/en/open_llm_leaderboard/about

McKinsey & Company (2025) The state of AI in 2025: Agents, innovation, and transformation. Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Meta via Hugging Face (2024) Llama-3.3-70B-Instruct model card. Available at: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

Microsoft via Hugging Face (2024) microsoft/phi-4 model card. Available at: https://huggingface.co/microsoft/phi-4

Minkkinen, M., Laine, J. and Mäntymäki, M. (2022) 'Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks', Digital Society, 1, p. 21.

Mistral AI (2025) Mistral Small 3 (24B; Apache-2.0), 30 January. Available at: https://mistral.ai/news/mistral-small-3

MIT Project NANDA (2025) The GenAI Divide: State of AI in Business 2025. Available at: https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Mitchell, M. et al. (2019) 'Model cards for model reporting', in FAccT Conference Proceedings, pp. 220–229.

Mökander, J. et al. (2024) 'Auditing large language models: A three-layered approach', AI and Ethics, 4(4), pp. 1085–1115.

NIST (2023) AI Risk Management Framework (AI RMF). Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf

OECD (2024) AI Principles. Available at: https://www.oecd.org/en/topics/ai-principles.html

Open Source Initiative (n.d.) Open Weights: Not quite what you've been told. Available at: https://opensource.org/ai/open-weights

Pulsora (2025) Enterprise guide to ESG reporting requirements in 2025, 8 September. Available at: https://www.pulsora.com/blog/enterprise-guide-to-esg-reporting-requirements

Qwen via Hugging Face (n.d.) Qwen2.5-7B-Instruct. Available at: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Raconteur (2025) Six green projects to make frontier AI more sustainable, 9 September. Available at: https://www.raconteur.net/technology/6-green-projects-to-make-frontier-ai-more-sustainable

Rein, D. et al. (2024) 'GPQA: A graduate-level google-proof Q&A benchmark', First Conference on Language Modeling.

Reuters (2025) AI models with systemic risks given pointers how to comply with EU AI rules, 18 July. Available at: https://www.reuters.com/sustainability/boards-policy-regulation/ai-models-with-systemic-risks-given-pointers-how-comply-with-eu-ai-rules-2025-07-18

Scaramuzza, F. et al. (2025) 'Show me you comply... without showing me anything: Zero-knowledge software auditing for AI-enabled systems', arXiv preprint arXiv:2510.26576.

Schröder, T. and Schulz, M. (2022) 'Monitoring machine learning models: A categorization of challenges and methods', Data Science and Management, 5(3), pp. 105–116.

Unit4 (2025) How can you prepare your organization for ESG reporting compliance with an integrated FP&A solution?, 9 May. Available at: https://www.unit4.com/blog/esg-reporting-will-soon-become-mandatory-heres-how-you-can-prepare-your-organization

United Nations Environment Programme (2025) New coalition aims to put artificial intelligence on a more sustainable path, 11 February. Available at: https://www.unep.org/news-and-stories/press-release/new-coalition-aims-put-artificial-intelligence-more-sustainable-path

Wang, Y. et al. (2024) 'MMLU-Pro: A more robust and challenging multi-task language understanding benchmark', NeurIPS, 37, pp. 95266–95290.

Yang, Z. et al. (2018) 'HotpotQA: A dataset for diverse, explainable multi-hop question answering', EMNLP. Available at: https://aclanthology.org/D18-1259/

Zheng, L. et al. (2023) 'Judging LLM-as-a-judge with MT-Bench and Chatbot Arena', NeurIPS, 36, pp. 46595–46623.