SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Designing Evaluation Frameworks For Enterprise LLM Deployments In Regulated Environments

Karan Nisar

University at Buffalo, Buffalo, New York knisar17@gmail.com

Abstract
Enterprise organizations are increasingly adopting large language models (LLMs) to support automation, decision-making, catering to customers, and managing knowledge. But using the technology in regulated areas like finance, health, insurance, and government raises significant concerns around compliance, privacy, equity, security, and accountability. Current evaluation approaches mostly emphasize technical aspects such as accuracy, fluency, and reasoning without mentioning enterprise readiness in the regulated environment. In this paper, we propose a framework to guide the evaluation of enterprise Large Language Models in regulated settings. The framework takes into account seven factors: accuracy, regulatory compliance, cybersecurity and privacy, bias, explainability and auditability, and enterprise readiness. The research suggests a weighted scorecard to inform the model of deployment, risk classification, and post-deployment monitoring. The study also investigates the contributions of governance methods such as human oversight, transparency, red-teaming, and post-deployment monitoring to enable safety and reliability. This approach bridges the gap between high-tech and enterprise risk management and offers enterprises a way to deploy scalable, safe Large Language Models that can be ready for enterprise. This study creates a governance approach to decision-making around safe, reliable, and scalable deployment of AI in critical domains.

Keywords: Enterprise AI, Large Language Models, Regulated Environments, AI Governance, Compliance Evaluation, Risk Management.

I. Introduction

Large Language Models (LLMs) are some of the biggest innovation breakthroughs of the digital age and are starting to impact how enterprises communicate with customers and employees, use enterprise information, and automate enterprise processes. They rely on neural networks with a transformer architecture that are trained on large collections of books, articles, web pages, code, enterprise data and enterprise knowledge bases such that they learn the statistical likelihood of language to generate natural language, summarize text and perform search over long text documents, answer questions from a large body of knowledge, classify information, generate computer code, translate languages, and now even train chatbots [1]. They are a more flexible form of artificial intelligence that has evolved from previous generations that were based on rule-based systems or supervised task-specific machine-learning models, and have some degree of generalized intelligence and language reasoning that can be tailored to different tasks through prompting, fine-tuning, Retrieval-Augmented Generation (RAG), and tool-access [2]. The fact that they are generalized means that they can be used to extend intelligence to systems, so enterprises can increase intelligence in systems without having to build new systems to perform the same tasks. These days, enterprise users are incorporating LLMs into customer-service systems, knowledge-sharing systems in enterprises, fraud-assessment systems, contract-analysis systems, human-resource support systems, research-summarization systems, marketing-content systems, and software-development assistants. These systems might be efficient; they can accelerate decision-making, make information more accessible, and work with humans. But there are issues with the move from "the lab" to "the line that cannot be overlooked [3]. Unlike controlled experimental environments where models are evaluated on curated benchmarks, real-world enterprise deployments expose LLMs to unpredictable inputs, shifting data distributions, domain-specific nuances, and stringent regulatory constraints. Challenges such as hallucination, inconsistent outputs, latency, cost at scale, data privacy, model explainability, and integration with legacy systems become critical concerns that must be systematically addressed. This gap underscores the urgent need for robust LLMOps frameworks, encompassing prompt versioning, output

validation, continuous monitoring, human-in-the-loop feedback mechanisms, and governance protocols, to ensure that LLM-powered enterprise systems are not only performant but also reliable, auditable, and aligned with organizational and ethical standards.
. The response might be probabilistic for the same question; might respond to questions that are difficult to prove; might not draw out potentially biased data; or might "leak" sensitive data if sample sets are not regulated. Other questions might relate to the speed of the response of the model, cost for model inferences, Service Level Agreements (SLAs) in terms of system availability, how the system integrates with enterprise IT systems, how the model is named and versioned, and its lifecycle. In the financial, insurance, health and drug discovery, telecommunications, energy, and government sectors, these issues are even more important as the model's response can be used to grant credit for loans, to recommend medical treatments, to determine insurance claims, or to issue compliance reports, or to provide public access to government services [4]. In these cases, the process for using the model should adhere to privacy, explainability, fairness, security, resilience, and auditability requirements. And so, the dynamic deployment of LLMs in enterprises cannot be determined simply by their efficacy in public evaluation or demonstrations. It needs more sophisticated ways to determine whether the model is technically competent and compliant (legal, ethical, risk), operationally sound, and strategically of value by contributing to enterprise risk management. Thus, developing those systems is part of the second wave of enterprise AI [5].

While there is a strong commercial momentum in the development and deployment of LLMs, the practice of evaluating LLMs is disparate, immature, and not aligned with enterprise needs. The majority of public benchmark ecosystems have been built to assess the quality of natural language-related research, rather than provide trustworthy enterprise practices. It's common to see a public leaderboard reporting the performance of LLMs with metrics such as BLEU, ROUGE, perplexity, MMLU, win-rate comparisons, and human preference, which is helpful to indicate the fluency, reasoning ability, or even programming ability, but does not satisfy enterprise concerns [6][7]. A system might have the highest score in the public leaderboard, but not be suitable for the needs of interpretability, privacy, identity and access control, cost, incident management, incident reporting, or regulation. The situation is that many firms are only experimenting with LLM tools on a case-by-case basis by different groups, which results in discrepancies with some cybersecurity, legal, data science, procurement, or business needs. One team may rate factors such as productivity, another team may be cybersecurity oriented (security mindset), or other teams may be cost-conscious (vendor costs, local data sovereignty) [7]. So far, market research suggests industry is now transitioning from experiments with LLM's to experimental deployments, and the question is now changing from innocuous: are they useful? to tricky: when and how can we trust them to be correct? Recent issues of hallucinated legal citations, privacy violations through prompts, prompt injection, biased answers, and performance degradation with fine-tuning exemplify dangers when only testing performance. And regulations are moving quickly [8]. The regulation of AI and data protection (e.g., EU GDPR) in the European Union (EU), financial regulatory guidelines, medical standards, and national AI trust schemes are bringing new aspects of risk category, risk management, explanation, and human control [9]. However, there remains a dearth of both academic and practitioner knowledge on integrated approaches on how to connect technical model verification, regulations, operational safety, and vendor risks, with business value. This is concerning, given that in some industries model performance faults can translate into regulatory, medical, legal, and reputation implications. So, the need is for a next-gen acquisition, trial, approval, monitoring, and management (by the board) of the full product development lifecycle of LLM systems [10].

This paper seeks to develop a holistic, business-relevant, and governance-focused approach to evaluating enterprise Large Language Models (LLM) deployments in compliance settings. Rather than limiting the scope of evaluation to model intelligence or benchmark integration, the paper provides an enterprise-operating-model view of deployment leading to success in achieving a fit between technology, governance, regulatory and legal considerations, operational resilience, and economic value. The aim of the framework is to serve decision makers such as chief information officers (CIOs), chief risk officers (CROs), compliance officers, data science teams, internal auditors, procurement officers, and business-unit sponsors who need to work together to decide whether adopting LLM is safe and valuable.

The focus of the paper is as follows:

- Multidimensional Evaluation Framework - Creating a framework of metrics for functional correctness, domain-specific semantics, rate of hallucination, security, privacy, fairness, explainability, regulatory compliance, and usability.
- Graded Readiness Maturity Models - Creating scorecard models that level LLM use cases as sandbox, pilot, controlled deployment, enterprise, or critical depending on thresholds for risk vs. performance.
- Industry-Specific Regulatory considerations - Considering evaluation criteria in different industries such as banking, health, insurance, pharmaceutical, telecom, and administration, as responsibilities vary.
- Adapted Post-Deployment Monitoring and Assurance - Implementing post-deployment monitoring and assurance controls based on drift detection, model re-evaluation, experience red-teaming, logging, incident reporting, and human-in-the-loop reviews and change management.
- Engineering Strategic Value Realization - Discerning how enterprise organizations can optimize speed of innovation with cost discipline for the token, productivity gains for employees, customer experience improvement, robustness, and return on investment.

Through these aspects, this paper helps bridge the perennial gap between experimentation and successful generative AI deployment in enterprises. It offers a framework for enterprises to make informed deployment choices, meet evaluation requirements, and sustainably deploy LLM use cases at enterprise scale under strict operational constraints. By doing so, this research elevates evaluation from a point-in-time technical assessment to an ongoing governance practice vital for first-class enterprise AI modernization.

II. Related Work

The recent development of enterprise use cases for Large Language Models (LLM) has attracted considerable attention in the research and practitioner community in the evaluation, governance, and ethical use. The initial work has primarily focused on enhancing language capability through scaling laws, benchmarking, reinforcement learning from human feedback (RLHF), fine-tuning, and so on [11]. The evolution of enterprise applications made researchers think about more systemic issues like hallucinations, biases, privacy, prompt attacks, and off-target [12]. On the other hand, studies in information systems and enterprise risk management also suggested that the impact of enterprise AI is not only linked to the model performance, but also control, accountability, and ultimately business performance [13-15]. In the financial and other regulated industries, research on model risk, algorithmic accountability, and trustworthy AI has also proposed ideas relevant to LLMs on transparency and auditability of models, independence of model development and deployment, and human-in-the-loop oversight [13]. Recent propositions by digital technology vendors and standards bodies have suggested merging these concepts into practices such as a code of conduct for ethical AI, cybersecurity, and governance control of the entire AI life cycle [14]. But these are still in the technical benchmarking and testing, compliance, cybersecurity readiness, or value for the company. Research may focus only on benign prompts or benchmarking testing, or more on ethics and less on practice. Likewise, policies & regulations establish conditions but not scoring on readiness for deployment. With industry moving from prototypes to scaling up deployment, there is increasing awareness in research of integrated methods of testing and benchmarking that include technical metrics, governance readiness, risk management, and regulatory compliance. This section looks at some of the progress so far in studies of technical benchmarking, responsible governance of AI, cybersecurity risk control, and value for enterprise operations. The survey reveals progress has been made, but research is needed on holistic scorecards for deployments of enterprise LLM in critical applications with high-stakes regulations [11][13].

2.1 Benchmarking and Technical Performance Evaluation

There are many papers on benchmarking the technical performance of LLMs with similar benchmark sets and leaderboards. Researchers developed metrics for language translation, reasoning, code generation, and information retention [12]. The benchmarks MMLU, BIG-bench, HumanEval, domain-specific, TruthfulQA, and others became the norm to compare vendors and model variants [11][12]. This work was useful to demonstrate the rapid improvements for zero-shot and few-shot learning, enabled with bigger models trained with more compute and data. Benchmarks also provided transparency, because they provided a way for enterprises to benchmark claims about performance by vendors. But there are some issues with benchmark papers in the enterprise. First, benchmarks include tasks that are non-natural and don't fit with enterprise needs, such as claims management, compliance, litigation summaries, or knowledge management. Second, benchmark results often don't reflect sensitivity to prompts, languages, and domain-specific terms. Third, public benchmarks may not measure variability and efficiency of response times, cost of tokens, and availability, integrability - important enterprise

metrics. Fourth, models may overfit to the benchmark data and result in a gross inflation of the performance metrics without usefulness [12]. Finally, tech benchmarks don't rank legal defensibility, compliance, and business value.

2.2 Responsible AI, Ethics, and Governance Frameworks

The second area of related work is on responsible governance, fairness, transparency, and ethics. This research grew rapidly as companies realized that leading technology models can pick up unfair biases, generate unfair outcomes, or recommend to users or stakeholders in inexplicable ways. Guidelines developed by standards bodies, universities, and multinational technology companies emphasize fairness, accountability, transparency, explainability, privacy, and contestability [13]. These are especially important for LLMs, as information may be misleading and seem more credible than other information. Enterprise research into governance includes policies, committees, independent model verification teams, and mechanisms to report model issues. Other studies advocate human-in-the-loop considerations for health advice, loan approvals, or hiring. Other lines of work promote model cards, documentation, and data lineage. These studies provide valuable institutional measures to reduce risk and build trust [13]. So, while research on responsible AI offers important normative advice, businesses need scorecards to operationalize governance norms. This work builds on these studies by harnessing ethical considerations and technical and business needs for model deployability [13][15].

2.3 Security, Privacy, and Adversarial Risk Research

A further body of relevant research is on how to secure enterprise systems for cybersecurity and privacy. As generative systems started gaining more traction, researchers began new vulnerability taxonomies for prompt injections, data extraction, jailbreak prompts, weak plugin-links, model inversions, adversarial fine-tuning, and third-party supply-chain [14]. These risks are more severe when scaling down to the enterprise level, as the model may access corporate documents, client data, code reviews, or other productivity tools. As such, the security literature proffers zero-trust, role-based access, output filtering, content moderation, logging, encryption, and segregation. Privacy research also pays attention to memorization, privacy, and retention, transboundary data flows, and accidental prompting. This is particularly relevant to privacy in the context of GDPR, sector-specific privacy in health and finance, and other confidentiality needs. Some approaches suggest hybrid retrieval-augmented models to hold private data outside of links, while others emphasize private hosting, secure enclaves, or federated learning [13][14]. As such, recent studies support integrated assurance approaches that include comprehensive testing of security, privacy, and technical performance [14]. This paper draws on this expertise by embedding cybersecurity and privacy as weighted elements in a value model for enterprise [14].

2.4 Business Value, Adoption Strategy, and Enterprise Transformation

Finally, another related literature stream is about the value of LLM. Case studies and consulting reports by management researchers reveal the use of LLMs by enterprises to reduce service costs, accelerate processes involving document review, improve worker productivity, lead to new modes of customer service, and enable the extraction of information and insights from unstructured data [15]. Examples include chatbots for support services, copilots for computer programmers, research assistants, drafting contracts, and translating languages. These studies show that the capacity to efficiently use generative AI could become a valuable capability. Many trials don't progress beyond experimentation because of diminishing returns on integration, governance, model accuracy, and/or user acceptance. For example, integration costs, vendor lock-in, employee training, and the need to change business processes can subdue outcomes. In addition, some organizations focus on innovation rather than delivering business value, leading to uncoordinated and strategic experiments [15]. This poses a key research question: evaluation tools need to provide information on which safe models to deploy and what safe deployments create value for business, given costs and governance issues. This research takes this into account by taking business value as an evaluation criterion, rather than measuring business value only after deployment [15].

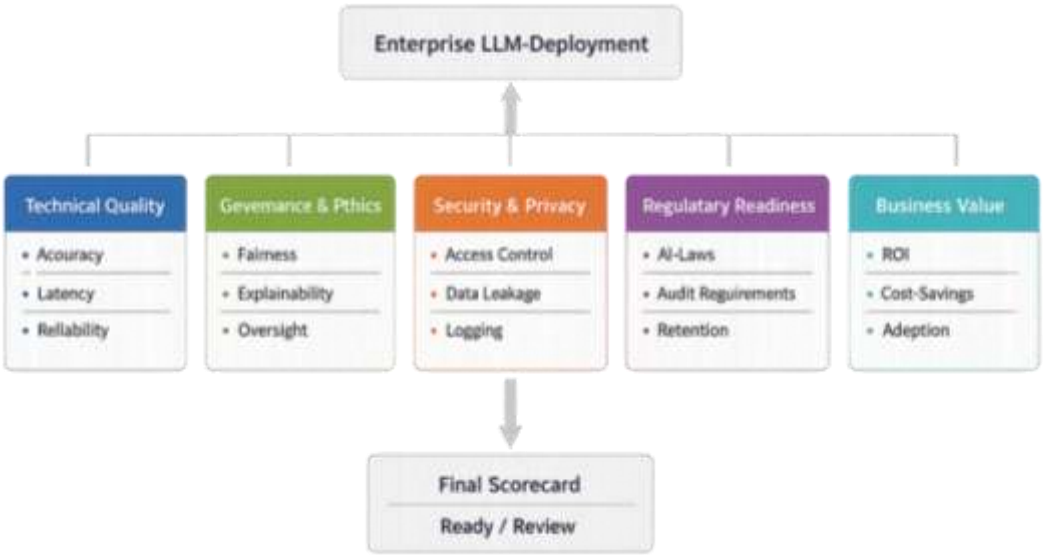
III. Proposed Enterprise LLM Evaluation Framework

We need a readiness model to accelerate the transition from proof of concept to an enterprise strategic capability for Large Language Models by measuring readiness from technical, operational, governance, and strategic perspectives [16]. Many innovators started to play with Large Language Models (LLMs) in their own silos for chatbots, summarization, code writing, knowledge, and customer service autoscaling applications. Many of these activities have created hype with quick wins in terms of speed, efficiency, and streamlining. But these were usually

primarily designed to support proof of concepts more so than enterprise use. As a result, they lacked appropriate security, audit, versioning, human-in-the-loop, data, and lifecycle management. Existing model validation and evaluation approaches, which concentrate on benchmark accuracy, language, or abstract reasoning skills, are not suitable for environments in which model output predictions may impact customers, patients, users, investors, regulators, or trust [17]. For example, a model that pleases the public may not be suitable if it cannot explain a decision, protect privacy, be secure for users, and comply with dataset retention periods. The challenge, therefore, in the enterprise adoption is a need for a broader definition of model intelligence - one that includes model deployability. In that vein, we explore a multi-faceted model for enterprise use cases, which is compliant with legal, security, and regulatory standards. Firstly, we consider the quality of the responses in terms of the accuracy of the response, consistency on repeated questions, on-topic, hallucinations, response time, throughput (tokens per second), multi-lingual or language proficiency, and aggregate task completion. They are used to establish the business-readiness of the model. The second layer looks at the governance quality measures of fairness, explainability, documentation quality, interpretation, human oversight, trust, accountability, and auditability [18]. That's particularly critical for financial, health, insurance, and government use cases. The third aspect relates to cybersecurity and privacy measures of prompt injection, access control, information leakage, encryption, secure integration, logging, and maturity, third-party vulnerabilities, and vendor risk [19]. A fourth measure concerns regulation and compliance, by understanding model use-cases and their implications on legislation, policies, norms, data retention (privacy), transparency, and reporting (local). Finally, data-centered measures of productivity, user experience, cost reduction, deployment and training, innovation support, revenue increase, and return-on-investment are taken into account to ensure the risk-effective deployment adds value [20]. Hence, the framework provides a win-win in a harmonized performance and governance risk-before-scaling.

Weights of metrics are adjusted to emphasize concerns in organizations, use cases, and risk profiles. For instance, private data, security, clinical understandability, and informativeness of documentation may be strained in clinical use, while auditability, bias testing, fraud, and decision traceability may be strained in financial use. Health insurance companies might be more interested in claims management and customer service fairness, and governments may be interested in inclusivity, transparency, trust, and fairness. When using a developer copilot, software companies may value system efficiency, security, and compatibility. That makes our framework flexible and fits the differences. Our model can be used during the procurement and deployment lifecycle process: procurement (evaluation and selection), pilot project approval, production (go/no go decision), procurer renewal, and monitoring. It is not a static process (certification) but defines an iterative governance process in which the model could be re-evaluated after a new version is released, after an incident, after a change in model architecture, a change of supplier, and/or a change of regulation. This is necessary as enterprise AI evolves with prompt tuning, retraining, API updates, and user updates. This helps the framework move from trial and error to information-driven deployment of enterprise LLMs with attributes such as reliability, trust, security, and alignment [16][17].

Figure 1. Proposed Multidimensional Evaluation Framework for Enterprise LLM Deployments



So, the importance of the proposed attribute is that it translates general concerns about AI governance into deployment considerations [17]. Even though many enterprises are aware of the use cases, benefits, and risks of using generative AI, they may not know if a particular model should be kept in the lab, undergo a pilot (or dry-run), or even go to production. In the absence of this, instead of identifying real risks, enterprises could rely on vendor promotions, bottlenecks, or technology demos. With weighted scores for the five areas, decision makers can prioritize vendors, prioritize different use cases (from within the firm), prioritize and aim investments to address the shortfalls in one or more risks, and gain insight into go or no-go in product release [18-20]. This is especially important if multiple models meet the technical criteria but vary in privacy, regulatory, governance, and/or maintenance costs. For example, a model may be technically good but non-compliant because its processes or operations have not been documented, do not have opportunities for human review or audit trail, or lack explainability. A different model may be compliant and secure but provide little value for the enterprise (thereby suggesting the enterprise should consider other, more valuable transactions). A third model may be more effective but less secure due to higher supplier risk, as this model is concentrated. In these various scenarios, the framework enables the enterprise to manage the effectiveness and validity of IT transactions. The framework also facilitates co-operation as it provides a framework for the risk managers, IT, procurement, legal, audit, and business leaders [19]. The stakeholders will no longer be speaking at cross purposes because their assessment of the readiness is based on their collective analyses of the evidence, threshold, and voting rules. This eliminates the tensions by answering the fast-track innovation developers (who want quickness) and the control teams (who want certainty). Gambits (and responses) can be reassessed as the law and the enterprise risk management policies change, or as model performance improves. This provides assured sustainability in an ever-evolving AI world. In highly regulated industries, where proper governance can potentially result in reputational, regulatory, legal, or consumer issues, litigation, customer churn, safety issues, or operational issues, the assessment is a key component of risk mitigation [18]. It also ensures oversight at the senior (C-suite and board) levels in terms of residual risk, control readiness, and value. Over time, roll-up scorecards can be used to benchmark between departments and for enterprise risk governance (ERG). Hence, we have an approach for institutional adoption of Large Language Models that is sustainable and ensures trust, compliance, resilience, and value creation [16][20].

IV. Implementation Of The Proposed Evaluation Framework

The deployment of the proposed enterprise LLMs evaluation framework shall be incremental, governance-oriented, and integrated, and link the technical deployment process with institutional risk, regulatory, cybersecurity, and business activities. The reason why it may not work in enterprises is that it is not a well-managed model deployment in enterprises. Consequently, the framework should be deployed as an enterprise program, rather than a technical project. The first step is use-case identification and classification, where enterprises determine if the intended LLM use case is for low-stakes tasks such as productivity apps, medium-stakes tasks such as decision support, or high-stakes tasks such as commercial use cases like support for lending, health communications, fraud reviews, claims, and citizen services. This is a key step, as it changes the focus of the evaluation criteria used, the level of approval authority required, the documentation needed, and the amount of review undertaken by the corporate governance process [21]. An example of this is that a corporate tool may require minimal controls, while a medical triage tool would require extensive review, human-in-the-loop, and privacy controls. The second step is data setup and integration, where the enterprise data is validated to support a secure LLM. This includes evaluating data sources, data privileges, identity, privacy, data integration, hosting, vendor service contracts, disaster recovery, and scalability. The enterprise must decide if the solution will be based on public APIs, private cloud, on-premise, sovereign cloud regions, and/or retrieval augmented systems. Regulations around data location, costs of tokens, response times, and seamless integration with existing systems must be considered. Improperly governed architecture decisions may lead to additional costs. Thus, implementation requires good collaboration between IT architecture, cybersecurity, and legal functions. The last phase is baseline model testing, where the final candidate models will be benchmarked for the technical aspects by using relevant prompts, hallucination detection, consistency, multilingual examples, adversarial prompts, throughput, and workflow completion rates. Each organization should be testing models on their own "real" prompts, not public benchmark corpora. A bank needs to look at the summarization of a complaint, description of a fraud or regulation, policies recall, not just academic reasoning. But technical success does not translate to OK to use. So, the fourth phase is of governance and assurance verification, such as checks for fairness, explainability, completeness of records, audit logs, cybersecurity, vendor due diligence process, and mapping to regulation or

industry standards [22]. Third-line-of-defense activities, such as internal audit, model risk management, or compliance committees, may be involved.

The fifth stage is experimentation with the use of the system by the business, with monitoring. Human supervision of the feedback on the ease of use, output, productivity, user acceptance, and effectiveness of controls occurs in this phase. The code suggests either rejecting, repairing, or deploying the system based on reported incidents and incident escalation, override, and satisfaction rates [23]. This is important as issues are often only discovered via user interactions. If the case is approved, the sixth stage is enterprise integration and lifecycle management, which includes integration of the LLM into processes, with access control, approval process, incident escalation, monitoring performance data and dashboards, drift, abuse of the prompts, cost, and periodic validations of the model. Periodic monitoring is needed as enterprise models, depending on regulatory, vendor models, data, or risks, can degrade over time. The enterprise should maintain governance records such as inventory and risk registers, vendor assessments, and audit trails. Finally, the implementation process should be owned by IT, legal, compliance, risk and cyber security, procurement, human resources (HR) training, and business stakeholders. Otherwise, technically advanced deployments of the LLM may not contribute to trusting and effective enterprise roll-outs. So, the model translates theory into practice to scale up the responsible AI enterprise [24][25].

V. Evaluation Scorecard Design

A readiness scorecard for the enterprise is one aspect of making the invisible governance needs visible to enterprise use of Large Language Models (LMs). In regulatory settings, enterprises can't simply think of the technology achieving the purpose, but also that the technology is safe and cost-effective, and predictably achieves the purpose in the enterprise's processes. The traditional measures of the system (e.g., accuracy relative to benchmark or speech quality) do not equate to readiness as they do not consider privacy and cybersecurity risk, readiness of documents, or investment planning. In contrast, the scorecard developed as part of this research is a weighted multi-criteria assessment based around these major components: Technical Quality, Governance and Ethics, Security and Privacy, Regulatory Readiness, and Business Value. This ensures readiness for use and readiness for control are factored in when enterprise decisions are made to use large language models. Technical Quality measures reasoning, consistency, hallucinations, and effectiveness based on benchmark tests (e.g., MMLU and HumanEval), and Governance and Ethics assess fairness, explainability, human oversight, and model accounting. Security and Privacy assesses prompt injection, RBAC (role-based access control), data breaches, and logging; Regulatory Readiness measures compliance with the NIST AI Risk Management framework and industry standards, etc. Business Value measures for productivity, adoption, cost, and CX. The weights of the different categories will vary with institutional incentives and use-case risk. So, a hospital will get more points for privacy and explainability than a bank, which will get points for non-bias and traceability. This enables the scorecard to be used as a framework for selecting, piloting, and rolling out AI vendors without the worst of the "Gartner effect", where the technology is more important than the governance [26][27][28].

Figure 2. Implementation Lifecycle of the Enterprise LLM Evaluation Framework

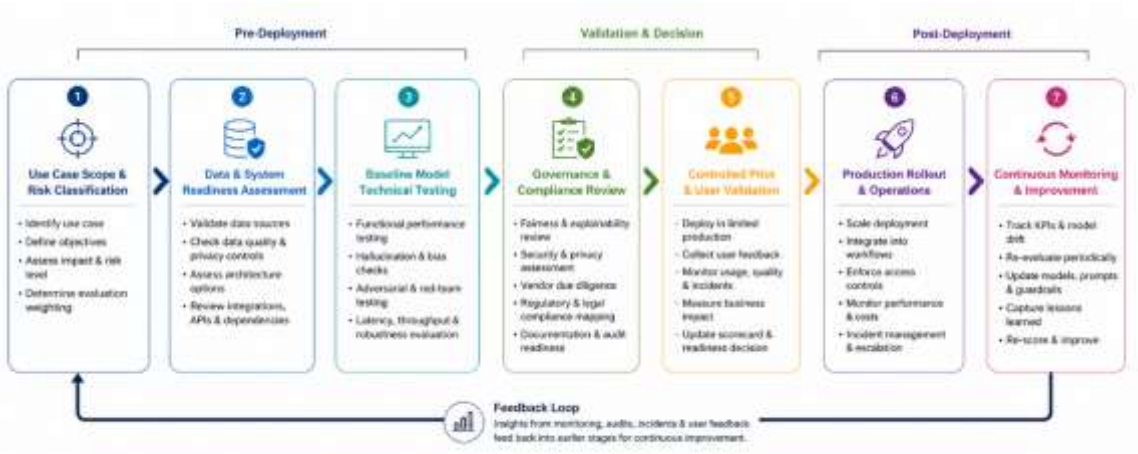


Figure 2. Implementation Lifecycle of the Enterprise LLM Evaluation Framework

Table 1. Enterprise LLM Weighted Evaluation Scorecard

Dimension	Indicator	Evidence from Original Sources	Weight (%)	Assessment Purpose
Technical Quality	MMLU Accuracy	GPT-4 reported 86.4% MMLU	25	Measures reasoning capability
Technical Quality	HumanEval pass@1	GPT-4 reported 67.0%	Included	Measures coding/problem-solving
Governance & Ethics	Human Preference Alignment	InstructGPT preferred over GPT-3	15	Measures safer/useful outputs
Security & Privacy	Prompt Injection Controls	OWASP lists it as a top risk	20	Measures attack resilience
Regulatory Readiness	Governance Lifecycle Controls	NIST Govern-Map-Measure-Manage	20	Measures compliance maturity
Business Value	Productivity Gain	~14% productivity increase in support tasks	20	Measures ROI potential

The true power of the scorecard is its implication in a readiness classifier model that helps specifically support enterprise governance and investment decisions, as well as deployments into control data rooms for the entire life cycle of Large Language Models. The readiness model is not just a technical degree, but rather a multi-parameter model of readiness assessment that is translated into an actionable model that can be applied by many stakeholders, such as the business, compliance, procurement, risk management, legal, and operational managers. Under this model, the five major aspects, Technical Quality, Governance and Ethics, Security and Privacy, Regulatory Readiness, and Business Value, are assessed separately, with common definitions, and then taken into consideration using weights such as business sensitivity, regulatory impact, operational criticality, and strategic importance. The model is then mapped to readiness labels such as Not Ready, Pilot Ready, Controlled Deployment Ready, or Enterprise Scale Ready, and in so doing, offers a safe journey from research to enterprise-scale. One of the benefits of such a classification is that it eliminates non-suitable and risky enthusiast-driven approvals. At many companies, the latest craze for state-of-the-art AI can sometimes trigger an itch to try out new tools using benchmark testing, vendor statements, or demos. Yet such performance doesn't equate to enterprise-readiness [29]. Even if a model performs at levels that exceed stringent benchmarking thresholds for reasoning capabilities, it might not be suitable for enterprise-scale deployment if it lacks other features, such as privacy, explainability, documentation, incident response, or compliance reports. Instead, a model with slightly lower benchmarks but more advanced governance, infrastructure, security, and compliance fit might be preferred for deployment in a production environment. This may be the case in sectors such as financial services, health care, insurance, pharmaceuticals, and government, where residual risks usually need to be documented and justified and meet approval to be deployed for production. In such industries, AI risks can result in loss or liability, harm to patients, reputation loss, fines, or lack of trust. The readiness model is also comparable between business lines/units, locations, and vendors by using the same decision rule across the organization, with location-specific differences in weighting if there are different legal/market environments. And it facilitates continuous improvement rather than a "one-time" pass/fail certification, since systems can be re-scored after cybersecurity improvements, retraining, user education, process redesign, policy, and/or vendor changes. Over time, enterprises can benchmark themselves on improvement, benchmark the vendors, identify common risks, and prioritize budget for addressing risks via readiness. In this regard, the scorecard is not just an evaluation method or technical standard; it's a governance practice that connects innovation with accountability, operational standards, executive oversight, and sustainable return on shareholder investment [30].

Table 2. Composite Readiness Classification Framework

Final Weighted Score	Classification	Recommended Action
0-49	Not Ready	Reject or redesign controls
50-64	Pilot Ready	Limited sandbox deployment
65-79	Controlled Deployment Ready	Department rollout with monitoring
80-100	Enterprise Scale Ready	Full production implementation

VI. Real Case Study Evaluation Outcomes

The effectiveness of the Large Language Model (LLM) conceptual enterprise toolkit scorecard notion here can only be evaluated in enterprise deployments where the benefits, constraints, and risks are also considered. Scorecard is normative in terms of key indicators, but enterprise deployments with AI systems indicate whether or not the business value is realizable for regulatory, technical, and human factors. Large Language Models (LLMs) for customer service, information management, document writing, computer programming, regulatory, and medical case management are being used in enterprises. Some LLMs have not been successful model deployments. While some have reported the benefits of speed, productivity, efficiency of human employees, and cost reduction, others have reported other issues, such as hallucination, privacy, lack of escalation mechanisms, fear of human workers on systems, and ill-defined governance. So, it could be that creating a sustainable enterprise market is more about good model deployment (through governance and readiness) than about the model itself. Therefore, scenarios (case studies) are key practices to illustrate the proposed deployment value. Studying their deployment instances (where balance is found) in the five elements of their scorecard: Technical Quality, Governance and Ethics, Security and Privacy, Regulatory Readiness, and Business Value, it becomes evident that there are wildly different enterprises with the same enterprise questions. A bank may place importance on privacy and auditability, and a hospital may not. An IT firm may value employee productivity and communication security, and a consulting firm may value reuse of corporate knowledge and confidentiality of client information. These showcase the maturity of the development of a single LLM capability. The following table illustrates examples of enterprise implementation of generative AI that have advanced beyond experimentation and development. Using the proposed readiness assessment, these are looked at for benefits, risks, and prioritization of elements for enterprise deployment. In conclusion, the case studies indicate that the enterprise deriving most value from LLM prioritized a use case, human management and control of the technology, and a person to own a deployment, and looked for value in the key performance indicators (KPI) from the deployment of LLM rather than bubble-generative AI as hype.

Table 3. Real Enterprise Case Studies Assessed Through the Proposed Evaluation Framework

Organization / Sector	LLM Deployment Use Case	Technical Quality Outcome	Governance / Compliance Considerations	Security / Privacy Factors	Business Value Achieved	Overall Readiness Interpretation
Morgan Stanley / Financial Services	Internal assistant for wealth advisors using proprietary research retrieval	High relevance in internal knowledge search and faster response generation	Strong need for audit trails, approved content boundaries, and advisory supervision.	Sensitive financial data requires strict access control	Improved advisor productivity and faster client preparation	High readiness when limited to internal advisory support

Klarna / FinTech	AI customer support assistant for handling service requests	Strong multilingual response capability and scalable support automation	Complaint handling, fairness, and escalation logic are essential	Customer data handling and authentication are critical	Lower service cost and faster response time	High readiness with continuous quality monitoring
Microsoft / Enterprise Software	Copilot tools for productivity software and coding assistance	Strong drafting, summarization, and coding productivity performance	Enterprise policy controls and responsible use standards are required	Data leakage prevention and tenant isolation are highly important	Large-scale productivity gains across enterprise users	Very high readiness in managed enterprise environments
Mayo Clinic / Healthcare	Administrative summarization and clinical documentation assistance	Useful for reducing manual paperwork and documentation burden	Human review is mandatory due to medical accuracy sensitivity	Patient privacy and health record confidentiality are critical	Improved clinician efficiency and reduced admin workload	Moderate to high readiness with strict oversight
PwC / Consulting	Internal knowledge assistant and proposal drafting support	Effective for document synthesis and internal search	Governance consistency is needed across global teams	Client confidentiality and document segregation are essential	Faster proposal cycles and staff efficiency	High readiness with strong internal controls
Duolingo / EdTech	AI tutoring and conversational language practice	High engagement and personalized learning interactions	Accuracy and fairness in learner feedback are important	User data privacy and minors' data controls are relevant	Greater engagement and a scalable tutoring model	High readiness in bounded educational workflows

VII. Sector-Specific Deployment Readiness Analysis

We can't use a "one size fits all" when it comes to readiness for deployment of enterprise Large Language Models (LLMs) because different industries may have different levels of regulatory oversight and risk of impacting operations, customer dependency, and risk. Many LLM systems have common features such as text generation and summarization, search assistance, and workflow automation, but the governance rules for deployment (and indeed usage) of the system can vary significantly between sectors. Enterprise LLM systems deemed ready for use for internal productivity gains by a tech company might not be ready for use in other sectors, such as healthcare and finance, without additional checks and balances. Consequently, deployment readiness for the particular sector for which the LLM will be used needs to be incorporated into the process proposed. In general, finance, government, health, insurance, and pharmaceutical industries will have a higher value for privacy, explainability, quality of documentation, human moderation, and auditability. By comparison, the retail, education, consulting, and enterprise software sectors could prioritize scalability, user experience, end-user productivity, integration, and security. The business sector readiness methodology allows organizations to establish a common governance approach, but adopt weights, thresholds, and process approvals as appropriate. It avoids the waste of extra time needed to govern less risky scenarios and scaled-down deployments in production. It allows multinational enterprises to have common templates while adhering to local sector regulations. The following sub-sections

briefly examine the needs of a particular sector and associated changes in weights of dimensions in the proposed readiness scorecard for that sector. This adds to enterprise AI governance by recognizing that it is not only about model performance when considering responsible deployment of AI, but also about impacts of failure, the sector perspective, and stakeholder trust.

7.1 Financial Services and Banking

The application of enterprise LLM technology in financial institutions is extremely delicate because the recommendations of the model could affect loan communication, fraud prevention, compliance and regulatory affairs, investment decisions, sanctions, anti-money laundering checks, and customer-facing operations. Banks should consider Regulatory Readiness, Security and Privacy, Explainability, and Documentation Quality in assessing readiness. Financial institutions need to explain their rationale, protect customer privacy, maintain documentation, and have robust internal controls. Despite being backroom tools, hallucinations or insufficient auditability of LLMs could present regulatory difficulties. Therefore, the best models might be used for search, summarization, and copiloting but not decision-making. Human supervision should be required for processes with customers. Models approved for use should have a super-high degree of maturity for scale.

7.2 Healthcare and Life Sciences

The medical industry needs to be highly risk-averse, given that our environment may affect communication with a patient, medical note-taking, coding, scheduling, and administrative efficiency. Even if the LLMs are not providing direct patient care, incorrect information or information that is not clear can result in safety concerns or administrative inefficiencies. Therefore, Privacy, Accuracy, Explainability, and Human Oversight are the most relevant factors to readiness. Organizations should ensure information privacy, integration with electronic health records (EHR), and escalation of emerging uncertain feedback. For life sciences/pharmaceuticals, other areas of concern are document integrity, compliance, and research privacy. LLMs excel at support with administration, taking notes, making summaries, and increasing productivity where humans make decisions.

7.3 Government and Public Administration

There are unique governance issues when using LLMs in government, as systems may be offered to citizens for direct use, support policy management, or be used to inform decision-making about services or access to services. In addition to usual considerations for Government concerning security and privacy, the critical considerations for government organizations are Transparency, Accessibility, Accountability, Language Inclusivity, and Procedural Consistency. Residents may expect some form of fairness for a government response, decision, or action, as well as understandability and the opportunity to appeal. So, readiness should be high. We recommend applications of the technologies in multilingual customer service, policy check, customer form assistance, and automation with appropriate handover to employees. Initially, fully automated decisions of high consequence should not occur.

7.4 Technology, Retail, and Professional Services

Less regulation applies to the companies providing telephone, internet, e-commerce, consultancy, and enterprise software service companies than to the financial and medical sectors, but there is skilled governance. For such companies, factors of Business Value, Productivity, Secure Integration, Usability, and Brand image are key readiness factors. Software development copilots, sales, proposals, knowledge chatbots, product insights, and complaint chatbots are frequently used. Massive adoption may take place, and hence, privacy, hallucinations to customers, shadow AI, and brand harm are still important aspects. These sectors may be implemented more quickly with a sequence of deployments (depending on user privilege, monitoring, and use controls). Readiness may be moderate-higher due to customer and data vulnerabilities.

VIII. Limitations And Challenges

Although enterprise Large Language Models (LLMs) are set to transform automation, decision-making, and efficiency in various industries, there are several limitations to adopting LLMs in a regulated environment. Although the proposed governance readiness assessment framework reduces uncertainty about the dynamic nature of LLM outputs, regulatory settings, and evolving cyber risks, there remain complexities and uncertainties that cannot be fully addressed. LLMs (unlike software) act more like words rather than commands, so their responses can vary depending on the prompt, context, and model, making consistency difficult. On top of this, businesses may not have the necessary technical capabilities in model validation, prompt engineering, and red-

teaming, and in managing the model lifecycle. These technical issues are compounded by vendor lock-in, lack of transparency regarding the use of training data, data sovereignty and cross-border considerations, as well as uncertain future costs associated with the number of tokens used and scaling requirements. And in regulated settings, we need to make generative AI solutions fair, explainable, and accountable, given that current regulations may not adequately mitigate risks. Despite the precautionary measures, humans are likely to assume trust and use the systems beyond their design. Further, in the rapidly evolving LLMs market, testing, security, and governance requirements may soon be out of date. This suggests enterprise readiness is a dynamic process or even a "badge of honor". The following subsections discuss the primary limitations and challenges of enterprise-scale deployment of LLMs across sectors in terms of technical, governance, integration, and economics. These constraints are important as the adoption of enterprise AI solutions requires an understanding of both the benefits and constraints that can limit trust, compliance, and added value.

8.1 Technical Reliability and Model Uncertainty

A problem with the technical risks posed by using LLMs for business is the consistency of responses. SysLLMs are probabilistic (they have variable responses, based on knowledge of the training data), unlike traditional computer systems, which are deterministic (they give the same response every time they are run, given a particular input). They may therefore alternatively respond in different languages, with longer or shorter prompts, or with different models. This introduces variability in applications, particularly in heavily regulated environments with expectations of consistent and predictable behavior. We should be concerned about hallucination, or strong but unrelated words, or nonsense. With enterprises, hallucinations may cause extraction mistakes, policy mistakes, wrong advice to customers, or flawed analyses. Failures may be infrequent, but a severe failure could lead to brand or regulatory issues. The second problem is the erosion in performance due to shifts in the needs of customers, understanding of business, or even modifications in regulatory policies. Failure to actively monitor could adversely affect long-term stability. And performance is also affected by adversarial extrapolations, language differences, and long-range comprehension. Models may perform well on English language-based tests, but be comparatively worse in local language or specific language models. They may also not perform well with big enterprise documents, for this reason. Thus, we cannot always rely on benchmarking of accuracy to be suitable for production. So, enterprises need to benchmark (and other methods) of control, supervision, and monitoring of a model when it is brought within the enterprise, rather than assume it is always accurate.

8.2 Governance, Compliance, and Accountability Complexity

LLM enterprise governance is the most difficult to implement as it is a business, legal, risk, and audit platform. No owners for enterprise AI system - so you need to establish "who is to blame" if the model is inaccurate and not compliant with enterprise policy, regulations, and complaints. This results in a lack of governance processes. Challenges in regulated industries. The enterprise may need to adhere to privacy regulations and laws, industry standards, model risk management policy, data retention policy, or other data policies. However, this may not be as transparent from LLM vendors in terms of the data used for the prompt, model, or safety forensics, and may be harder to verify. This can become a trade-off between transparency and the assurances that the enterprise will get. And there's explainability. Some really useful work needs explanations of what's done or what's meant. Complex networks are not very explainable. A user's activities (actions and searches) are useful for monitoring, but not explaining. It's also difficult to achieve mathematical equality as cultures, languages, or users can be discriminatory. The guidelines are not comprehensive, which can be hard for businesses. So, there are costs for employees, the board, and professional audits.

8.3 Operational Integration and Workforce Adoption

Companies overlook the effort required for system integration with a model, databases, identity managers, approval processes, UX, etc. Without achieving full system integration, employees' option for unauthorized methods creates shadow AI security issues. The largest challenge lies in redesigning work processes. LLMs change work processes through new ways of doing tasks. Customer service agents do not build responses but edit LLM responses. Analysts review summaries created by AI, rather than locating documents. Managers need to establish new control mechanisms and redesign jobs so as to have new performance appraisals. Without redesigning processes, productivity will only slightly increase. Users' trust is crucial. Users either trust the system too much and don't verify the system output, or some users don't trust the system model and don't use the system. Trust and distrust have a negative effect on an organization. There needs to be training provided to employees that instructs them on when to use AI, when not to use AI, when to challenge AI, and to whom to go for support. Organizations

need change management plans, which are critical for them to collaborate in large teams that communicate using different languages and have different levels of digital literacy. Organizations must have communications that include training depending on their roles in the organization, feedback loops, and leadership support. LLM implementation will not be effective until organizations upskill their employees

8.4 Economic Sustainability and Vendor Dependency

Companies that begin to incorporate LLM technologies because they believe the technologies will provide rapid economic returns will find it hard to remain economically viable. Ownership costs will rise due to the need to pay for token pricing data and API calls, to scale up infrastructure and fine-tune and monitor the models, and to provide governance personnel. The initial small-scale pilot, which seems to be cheap, will become costly when it must be scaled up to support thousands of staff members and millions of interactions with customers. Companies struggle to monitor return on investment figures as they are unable to put in place reliable measurement procedures. There will be enhancements in productivity, which will occur due to different reasons in different departments. Some users will reap considerable benefits, and others will see small gains. The calculated cost reduction will be undermined by increased review work, the costs of cybersecurity, and the need to change workflows. The team needs to manage executives' expectations. The company has another strategic challenge due to its dependency on vendors. Organizations rely on a limited and select group of external model providers who provide API and hosting, and system functions. Organizations are threatened by two issues because of their excessive reliance on external model providers who offer essential system functions. Organizations need to develop their systems to suit their business needs, as it will be too expensive to change their suppliers if their systems are integrated with proprietary systems. Organizations need to work on new architectural techniques because, in the near future, actual technological leadership will be revealed due to market dynamics. The investments of enterprises for the LLM should be managed as flexible investment portfolios to provide them with full business freedom. Organizations ensure continuous value by following three key principles: they are highly cost-conscious, develop multivendor relationships, and assess their business effectiveness.

IX. Future Research Directions

The unprecedented speed of enterprise Large Language Models (LLMs) development has a pressing need for future research to build on the current enterprise deployment architecture. This analysis brought forward an approach to evaluate enterprise systems for the regulated industry, but enterprise deployment settings in the future will be dictated by advances in multimodal AI, smart agents, industry-specific foundation models, trustworthy AI, and changing regulations. As enterprises transition from thinking of LLMs as copilots to agents executives, system evaluation must transition from monitoring performance to creating accountability for decisions, managing risks, and dependability in orchestrating across systems. The scorecards work for today's systems, but future systems may be continuously evolving, use tools and orchestrate other agents, and generate actions and recommendations instead, or in addition to text. This raises new challenges in terms of controllability, explainability, escalation policies, liability, and dynamic governance. Furthermore, there will be a need for standardized assurance models to support across-border operations with legal, privacy, industry, and other regulatory considerations. Economic issues will also come into focus as companies seek to control cryptocurrency token use, compute infrastructure spending, and greenhouse computing energy usage in addition to managing business relationships across multiple suppliers. The second of the research priorities relates to human factors, especially trust, over-reliance, skills decay, and job redesign for industrial AI-enhanced systems. Finally, measurement science will need to catch up as fixed benchmark ratings will no longer apply in a reality where AI models are in constant flux and interact with real-time enterprise data feeds. Research needs to consider new governance with adaptive, evidence-based, and industry-specific insights to support the safe scaling of enterprise AI systems. We provide some discussion on four areas of research: self-assured AI, privacy-preserving enterprise architecture, benchmarking in situ, and the role of work and process adaptation.

9.1 Autonomous Agents and Action-Based AI Governance

The main objective for future designs of the language support systems is to build task-executing agents. Existing enterprise LLM systems only work in consultative mode due to a requirement to finalize decisions. The system designs now support users to book meetings, process claims, generate reports, initiate transactions, and access databases as they conduct complex business processes. The advancement of the system poses a challenge for decision-making because it requires different techniques to respond to the new system. Researchers must find

what evaluation system should be used to evaluate the new system that takes actions, rather than the system that gives answers. The old system that evaluates the correctness of the answer and language proficiency no longer fulfills the need because the new models have the capabilities to perform financial actions, make changes in clients' records, and send emails to clients. The new ways to evaluate will define what actions are allowed, how the actions are rolled back and escalated, and how system actions are monitored and what is done in case of system failure. Liability will be an issue when the AI agent executes unethical and non-compliant actions because there will be multiple parties at fault, including vendors, developers, operators, and business owners. Future research should explore multi-agent systems that collaborate across departments. It will cause errors when several models work together because each model will work separately without errors. The governance system should monitor three key risks, namely coordination risk, conflicting goals, and cascading failures. Companies need evaluation methods that operate more like production control to deliver the desired outcomes and not the traditional model testing and evaluation approaches.

9.2 Privacy-Preserving and Sovereign AI Architectures

The research needed for enterprises that leverage generative AI systems for their business needs will focus on privacy and data rights management. Current systems today rely on the use of third-party application programming interfaces and stored foundational models (FMs) in public cloud computing environments, which pose challenges for managing the sensitive data of their customers and internal confidential documents, and for managing the business and trade secrets and cross-border data transfer considerations. Governed industries need to adopt enhanced measures for managing their data storage and processing operations and retention policies. Researchers need to research privacy protection systems that involve deploying the LLM locally in the premises of the organization, isolated secure computing, decentralized learning, confidential computing, data-restricted systems, retrieval-augmented generation, and the development of synthetic data for tuning the LLM. Researchers need to run comparative studies that will help to identify differences in how different privacy protection measures impact different performance outcomes, together with increased costs of infrastructure and complexities of execution. The installation of all-privacy systems will enhance trust in the corporate governance model, but it will increase costs while limiting access to the latest advances in model development. The development of sovereign AI infrastructure will enable nations and sectors to build national control systems that will establish their own AI models that adhere to local regulations and language use. Data protection will become essential for the health, defense, financial, and regulatory systems.

9.3 Sector-Specific Benchmarks and Continuous Evaluation Science

Evaluation test benchmarks currently allow researchers to test models and model performance, but future enterprise research needs to test and develop enterprise-specific test benchmarks, which are constantly evolving in science. Public leaderboards mainly test abilities of reasoning, code, and summarization, understanding of academic knowledge, but they are not benchmarks of enterprise business processes. For example, a hospital may require a safe summary for discharge, an insurance company may require a fair explanation for the claim, and the bank may require a discharge summary. Hospital, insurance, and bank benchmark suites, which access business everyday workflows using industry-specific vocabulary and require multi-lingual capability and privacy of sensitive information, should be established in the future. The benchmark should cover two tests: test assessment accuracy, the power of accurate information search and execution, the efficiency of document management, and the outcomes on users' convenience. Specialized benchmarks will allow for benchmarking on operational-related aspects. What the current arrangements are doing to organizations is that they are now having to make constant changes to their models. Once benchmark reports have been generated, they may be old news in months or weeks. It may be time for the academic community to consider the continuing circle of testing models with current policies, threat/risk data, and operational environments. This could involve the use of virtual spaces, shadow testing, and red-teaming and re-scoring

9.4 Workforce Transformation and Enterprise Operating Models

The long-term success of businesses that use LLM technology relies on two things: technological achievement and the company's ability to change business processes. Future work needs to focus on how firms reshape their employees with the adoption of AI technologies. While past studies primarily focus on how AI technologies enhance productivity, they do not explore what occurs to the tasks when AI technologies generate documents, meetings, computer code, and other tasks for users. As unique tasks emerge, workers will undertake new strategic tasks, which will reduce the need for routine work. The management needs to develop new ways to train new and reskill

active workers, introduce systems for regulating productivity, and reward employees. Studies need to be conducted on interview positions that can be supported by AI and responsibilities these jobs need to retain, while companies preserve the skills of employees through the AI system output. Leaners often have too much trust in AI recommendations, or they do not use a potential system because they don't trust its value. The research should explore users' decision-making, how they employ interface design elements and transparency features, combined with skills. Companies need new management models as AI governance committees and model risk and cross-area oversight groups are part of the business operations system. The research should determine whether a central vs. decentralized AI system through business units provides better performance. The enterprise AI operating model is important for performance, such as model selection.

Conclusion

The rapid advancements in Large Language Models (LLMs) present a compelling opportunity for enterprises to boost productivity, offer faster support for decision-making, delight customers, and automate workflows. But the path to enterprise deployment from proof-of-concept is not as straightforward in regulated environments, with a high degree of privacy, fairness, explainability, cybersecurity, and legal accountability expected. Our study found that existing approaches that solely focus on benchmark performance and/or language fluency are insufficient to evaluate for deployment. Instead, enterprises require more robust governance frameworks that take into account system quality, but also governance, regulatory, security, and business-readiness issues across a system's lifetime. The proposed governance model included a scoring of readiness across five key aspects: Technical Quality, Governance and Ethics, Security and Privacy, Regulatory, and Business Value. Using weighted measures and readiness levels helps enterprises make better decisions about their pilots, partner assessments, roll-out, and monitoring. The paper also highlighted that not all industries have the same deployment priority, specifically, financial, medical, insurance, government, and technology. With this, we need to ensure we have an adaptive scoring system. Through the case studies, we also noted that deployments that work are those that include the development of narrow use cases, monitoring, security, and evaluation of use cases. Other challenges include model uncertainty, dynamic regulatory environment, integration issues, employee adoption, and maintenance costs. Do these challenges suggest that enterprise governance of Large Language Models needs to be a moving, rather than a static, exercise? Most importantly, adoption is speedy and dependent on striking a balance between control and innovation. Companies that focus on thorough evaluation, constant evaluation, and integration in AI initiatives are likely to ensure they responsibly use Large Language Models while ensuring trust and stability, and gaining a competitive advantage.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
[2] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
[3] Ivković, J., & Ivković, J. L. (2023). Conceptual analysis and potential applications of DL NN transformer and GPT artificial intelligence models for the transformation and enhancement of enterprise management, EIS/ESS, and decision support integrated information systems. EIS/ESS and Decision Support Integrated Information Systems.
[4] Chui, M., Roberts, R., & Yee, L. (2022). Generative AI is here: How tools like ChatGPT could change your business. Quantum Black AI by McKinsey, 20, 1-5.
[5] Act, A. I. (2024). Artificial Intelligence Act. Artificial Intelligence Act. eu,(Accessed: 28 February 2025), https://artificialintelligenceact. eu.
[6] RISK, C. (2021). Risk management framework. CORPORATE REPORT.
[7] Ashurbayev, O. (2025). LEADERSHIP'S ROLE IN AI ADOPTION AND DIGITAL TRANSFORMATION: BUILDING TRUST AND TRANSPARENCY FOR SUSTAINABLE INNOVATION. Latin American journal of education, 5(6), 257-277.
[8] Abdulqader, Z., Abdulqader, D. M., Ahmed, O. M., Ismael, H. R., Ahmed, S. H., & Haji, L. (2024). A responsible AI development for sustainable enterprises: A review of integrating ethical AI with IoT and enterprise systems. Journal of Information Technology and Informatics, 3(2), 129-156.
[9] Hadfield, G. K., & Clark, J. (2023). Regulatory markets: The future of AI governance. arXiv preprint arXiv:2304.04914.

[10] Whitfield, R. (2025). The Global AI Governance Alliance. Federalist Debate (16829670), 38(3).
[11] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.
[12] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. X. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research.
[13] Ward, S. (2005). Risk Management: organization and context. Witherby Insurance & Legal.
[14] Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., ... & Li, Q. (2024). Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. arXiv preprint arXiv:2401.05778.
[15] Fasha, M., Rub, F. A., Matar, N., Sowan, B., Al Khaldy, M., & Barham, H. (2024, February). Mitigating the OWASP Top 10 for large language model applications using intelligent agents. In 2024, the 2nd International Conference on Cyber Resilience (ICCR) (pp. 1-9). IEEE.
[16] Gueorguiev, T. (2025). An approach to integrate Artificial Intelligence in ISO 9001-based quality management systems. Measurement: Sensors, 38, 101787.
[17] Benraouane, S. A. (2024). AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems. Productivity Press.
[18] Gueorguiev, T. (2024, June). The Process Approach in Artificial Intelligence Management Systems. In 2024, the 9th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE) (pp. 1-4). IEEE.
[19] Jonnakuti, S. (2024). LLMs in the cloud: Best practices for scaling generative AI in regulated industries. International Journal of Engineering Technology Research & Management, 8(2).
[20] Ettinger, A. (2025). Enterprise architecture as a dynamic capability for scalable and sustainable generative AI adoption: Bridging innovation and governance in large organizations. arXiv preprint arXiv:2505.06326.
[21] Parimi, S. K., & Yallavula, R. (2025). Generative AI for Enterprise Trust: A Governance-Aligned Framework for Safe and Transparent Automation at Global Scale. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 6(1), 218-225.
[22] Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 1112-1123).
[23] Zhang, W. (2025). Operationalizing Generative AI: A Comprehensive Framework for LLMOps Excellence. Nvpubhouse Library for European International Journal of Multidisciplinary Research and Management Studies, 5(11), 82-89.
[24] De Almeida, P. G. R., Dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: a framework for governance. Ethics and Information Technology, 23(3), 505-525.
[25] Kapusta, A. S., Jin, D., Teague, P. M., Houston, R., Elliott, J., Park, G. Y., & Holdren, S. S. (2025, May). A framework for the assurance of AI-enabled systems. In Assurance and Security for AI-enabled Systems 2025 (Vol. 13476, pp. 109-120). SPIE.
[26] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
[27] Gallifant, J., Fiske, A., Levites Strekalova, Y. A., Osorio-Valencia, J. S., Parke, R., Mwavu, R., ... & Pierce, R. (2024). Peer review of GPT-4 technical report and systems card. PLOS digital health, 3(1), e0000417.
[28] Chen, B., Chen, K., Hassani, S., Yang, Y., Amyot, D., Lessard, L., ... & Varró, D. (2023, September). On the use of GPT-4 for creating goal models: an exploratory study. In 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW) (pp. 262-271). IEEE.
[29] Schuett, J. (2024). Risk management in the artificial intelligence act. European Journal of Risk Regulation, 15(2), 367-385.
[30] Al Naqbi, H., Bahroun, Z., & Ahmed, V. (2024). Enhancing work productivity through generative artificial intelligence: A comprehensive literature review. Sustainability, 16(3), 1166.