The American Journal of Interdisciplinary Innovations and Research
ISSN 2642-7478 | Open Access

Type Original Research
PAGE NO. 92-98

Check for updates

An Integrated MLOps Framework for Robust, Scalable Deployment of Large Language Models Across Domains

Dr. Tobias Müller
Institute for Renewable Energy Systems, Technical University of Munich, Germany

OPEN ACCESS

SUBMITTED 01 November 2025
ACCEPTED 15 November 2025
PUBLISHED 30 November 2025
VOLUME Vol.07 Issue 11 2025

CITATION
Dr. Tobias Müller. (2025). An Integrated MLOps Framework for Robust, Scalable Deployment of Large Language Models Across Domains. The American Journal of Interdisciplinary Innovations and Research, 7(11), 92–98. Retrieved from https://theamericanjournals.com/index.php/tajiir/article/view/7076

Abstract- The rapid proliferation and success of large language models (LLMs) across domains — from natural language processing, finance, medicine to multimodal tasks — highlight their transformative potential for research, industrial applications, and societal impact. However, scaling LLM deployment in real-world, production-grade environments introduces significant challenges in reproducibility, maintainability, performance optimization, and quality assurance. This article proposes a comprehensive, conceptual MLOps-centric framework for the deployment and lifecycle management of LLMs, integrating continuous integration/continuous delivery (CI/CD) pipelines in cloud-based settings, combined with domain-aware evaluation and governance strategies. Drawing on extensive literature — including surveys of LLM architectures and applications (Minaee et al., 2024; Naveed et al., 2023; Pahune & Chandrasekharan, 2023; Zhao et al., 2023), domain-specific use cases in finance (Lee et al., 2024), medicine (Gao et al., 2023; Dada et al., 2024), information retrieval (Zhu et al., 2023), multilingual models (Yuan et al., 2023), and multimodal expansions (Wang et al., 2024) — as well as recent work on MLOps practices and tool ecosystems (Chandra, 2025; Berberi et al., 2025; Zarour et al., 2025; Kazmierczak et al., 2024), we articulate the architectural components, workflow stages, evaluation metrics, risk-mitigation strategies, and domain-adaptive customization necessary for sustainable deployment. We discuss in detail the benefits, limitations, and future directions, including adaptability for specialized

domains, governance, reproducibility, and cross-domain interoperability. Our framework aspires to serve as a reference blueprint for researchers, engineers, and stakeholders seeking to operationalize LLMs effectively and responsibly.

Keywords: Large Language Models, MLOps, CI/CD pipelines, model deployment, domain adaptation, scalable AI infrastructure

Introduction

The advent of large language models (LLMs) has revolutionized the landscape of artificial intelligence and natural language processing. Models leveraging large transformer-based architectures — for instance, models inspired by BERT or GPT — have demonstrated impressive capabilities in natural language understanding, generation, translation, summarization, and even code generation (John Snow Labs, 2024; Minaee et al., 2024). The scale and flexibility of LLMs have enabled their adaptation to multiple domains: from finance and risk modeling (Lee et al., 2024), to healthcare and medical research (Gao et al., 2023; Dada et al., 2024), to cross-lingual applications (Yuan et al., 2023), and multimodal tasks that include vision as well as language (Wang et al., 2024). The transformative potential of LLMs across these domains is widely acknowledged, but realizing this potential in real-world, production-grade systems imposes substantial engineering, operational, and governance challenges.

Traditional machine learning development workflows — often ad hoc and iterative — do not scale well when dealing with LLMs, especially at enterprise scale. The computational demands, model complexity, necessity for frequent updates, and domain-specific evaluation criteria call for robust operational practices akin to software engineering. In recent years, the concept of MLOps — a combination of machine learning (ML) and DevOps — has emerged to address these challenges. MLOps introduces structured workflows, tooling, automation, monitoring, and governance tailored to machine learning systems (Berberi et al., 2025; Zarour et al., 2025; Kazmierczak et al., 2024).

Furthermore, recent empirical and conceptual studies demonstrate the need for integrating CI/CD pipelines, cloud infrastructure, model versioning, continuous training, and domain-specific evaluation models to support LLM deployment (Chandra, 2025). However, a gap remains in the literature: while there exist surveys on LLM architectures and applications, and separate discussions on MLOps practices and platforms, there is no comprehensive, domain-agnostic, end-to-end framework combining LLM-specific demands with MLOps best practices. The literature lacks an integrated blueprint that systematically maps how CI/CD pipelines and MLOps tooling can support LLM lifecycle management across diverse domains, ensuring reproducibility, performance, compliance, and scalability.

This article seeks to bridge this gap. We present a comprehensive, conceptual framework for the deployment and lifecycle management of LLMs, grounded in existing literature. Our aim is not to introduce new empirical results, but to synthesize extant knowledge into a coherent, actionable blueprint that can guide practitioners and researchers in deploying LLMs responsibly and efficiently. The framework addresses infrastructural components (cloud-based environments, CI/CD pipelines), governance (version control, evaluation, compliance), domain-specific adaptation and evaluation (e.g., finance, medicine, multilingual, multimodal), and operational scalability. We believe such a framework is timely and necessary, given the accelerating adoption of LLMs in industry and academia, and the complexity of deploying them correctly.

Methodology

Given the goal is to propose a conceptual, integrative framework rather than to report new empirical experiments, our methodology consists of a comprehensive literature synthesis and mapping exercise. We systematically review scholarly articles, surveys, and industrial-level white papers focusing on two primary axes: (1) LLM features, architectures, domains, and performance characteristics; (2) MLOps practices, CI/CD pipelines, and production deployment challenges. Our sources include peer-reviewed journals, preprints, surveys, and reputable industry publications.

Selection of Literature

We began with foundational surveys summarizing the architectures, capabilities, and trends of LLMs, such as (Minaee et al., 2024), (Naveed et al., 2023), (Pahune & Chandrasekharan, 2023), and (Zhao et al., 2023). These works provide a comprehensive view of the design space of LLMs, including architecture variants, pretraining strategies, fine-tuning approaches, and domain

adaptation. We then incorporated domain-specific studies highlighting use cases and challenges of LLM deployment: (Lee et ., 2024) for finance, (Gao et al., 2023) and (Dada et al., 2024) for medicine and clinical language understanding, (Yuan et al., 2023) for multilingual LLM evaluation, and (Wang et al., 2024) for multimodal expansions including vision + language.

On the operational side, we integrated literature from the emerging field of MLOps: recent systematic reviews address platforms, tools, best practices, maturity models, and challenges (Berberi et al., 2025; Zarour et al., 2025). We also leverage targeted technical discussions on continuous delivery and automation pipelines in machine learning contexts (Kazmierczak et al., 2024), as well as a recent article demonstrating CI/CD-based optimization of LLM performance in cloud environments (Chandra, 2025).

Synthesis and Framework Construction

We conducted thematic coding of the collected literature, extracting recurrent themes, challenges, and proposed solutions. These themes were grouped along axes such as computational infrastructure, versioning and reproducibility, continuous training and delivery, domain-specific evaluation, monitoring and governance, and adaptability across domains. Based on this thematic mapping, we developed a conceptual architecture — a modular, layered framework — which defines the key components and workflows needed for sustainable LLM deployment.

In addition, we elaborated potential evaluation and governance strategies for LLM deployment, drawing from domain-specific evaluation protocols (e.g., clinical evaluation benchmarks from (Dada et al., 2024), financial risk evaluation considerations from (Lee et al., 2024), multilingual fairness and coverage metrics from (Yuan et al., 2023)). We considered operation-level metrics (latency, throughput, resource utilization), model-level metrics (accuracy, robustness, bias), and process-level metrics (pipeline reproducibility, deployment frequency).

Finally, we conducted a critical analysis — evaluating advantages, potential risks, limitations, and future paths — to provide a balanced, research-grade discourse that acknowledges current knowledge gaps and open challenges.

Results

From our analysis, we distill the following main observations and propose the corresponding elements of the integrated framework.

1. LLM Diversity and Domain Variation

- The literature confirms a wide diversity in LLM architectures, pretraining strategies, and domain-specific adaptations. Surveys by (Minaee et al., 2024) and (Naveed et al., 2023) document transformer-based LLMs optimized for general language modeling, whereas domain-adapted versions specialize for medical, financial, multilingual, or multimodal tasks. (Pahune & Chandrasekharan, 2023) and (Zhao et al., 2023) categorize LLMs into several classes depending on training data, architecture size, fine-tuning or instruction-tuning, and multicast or single-task orientation.

- Domain-specific studies highlight that LLM performance and evaluation criteria vary substantially across domains. For example, (Lee et al., 2024) emphasize regulatory compliance, financial risk, interpretability, and auditability for finance applications; (Gao et al., 2023) and (Dada et al., 2024) focus on reliability, correctness, safety, and ethical considerations for medical and clinical use. Multilingual considerations bring additional challenges in coverage and fairness (Yuan et al., 2023). Multimodal tasks, such as combining vision and language (Wang et al., 2024), require architectural and evaluation adaptations beyond text.

- These observations imply that any unified deployment framework must support domain-specific customization, evaluation, and governance.

2. Need for MLOps and CI/CD for Scalable Deployment

- Traditional ML pipelines are not sufficient for LLMs due to model scale, computational demands, frequent updates, and complex evaluation needs. The MLOps literature underscores this: (Berberi et al., 2025) outlines a landscape of platforms and tools; (Zarour et al., 2025) surveys best practices, challenges, and maturity models; (Kazmierczak et al., 2024) elaborates on continuous delivery and automation pipelines tailored to ML.

- Specifically for LLMs, (Chandra, 2025) demonstrates how CI/CD pipelines in cloud-based environments can optimize performance, streamline update cycles, and improve reproducibility. The analysis reveals gains in consistent deployment, automated testing, performance monitoring, and streamlined collaboration

among developers and researchers.

○ Nevertheless, across examined literature, there is an absence of a standardized, domain-agnostic blueprint combining these practices with domain-specific evaluation criteria and governance.

3. Critical Role of Domain-Aware Evaluation and Governance

○ Domain-agnostic evaluation metrics (e.g., perplexity, BLEU, validation loss) are inadequate for production contexts. Domain-specific evaluation must consider regulatory compliance, fairness, interpretability, safety, and user-centered criteria. Studies in medicine (Dada et al., 2024), multilingual tasks (Yuan et al., 2023), finance (Lee et al., 2024) and multimodal vision-language tasks (Wang et al., 2024) highlight these domain-specific requirements.

○ For example, in clinical applications, evaluation frameworks like the CLUE benchmark (Dada et al., 2024) are needed to test clinical language understanding, safety, and adherence to medical ethics. In finance, evaluation must consider risk, transparency, compliance, and explainability (Lee et al., 2024). In multilingual applications, coverage, fairness across languages, and bias must be measured (Yuan et al., 2023).

○ This suggests that LLM deployment frameworks must embed domain-aware evaluation protocols and governance mechanisms rather than rely solely on generic metrics.

4. Proposed Integrative Framework: Modular, Layered Design

Based on the observations, we propose a modular layered framework consisting of the following layers:

○ Infrastructure Layer: cloud-based compute resources (GPUs/TPUs), storage (object storage for model artifacts, dataset repositories), orchestration systems (containers, Kubernetes, serverless), scalable resource allocation, and hardware abstraction.

○ Versioning and Model Registry Layer: a registry for model metadata (version, configuration, training data provenance), dataset versioning, checkpointing, hashing, provenance tracking, and metadata about fine-tuning or instruction-tuning.

○ CI/CD Pipeline Layer: automated pipelines that handle code changes, data versioning, automated training or fine-tuning, automated testing, automated evaluation (unit tests, integration tests, evaluation tests), automated deployment to staging and production, rollback mechanisms, and integration with monitoring.

○ Evaluation and Validation Layer: domain-specific evaluation suites including benchmark tests, fairness and bias assessments, compliance audits, security, adversarial robustness checks (especially for safety-critical domains), latency and performance tests, resource utilization profiling.

○ Monitoring and Logging Layer: runtime monitoring (throughput, latency, errors), logging user queries and responses (with privacy safeguards), usage statistics, feedback loops, drift detection (data/model decay), anomaly detection, continuous performance tracking.

○ Governance and Compliance Layer: version governance, access control, audit trails, approval workflows for deployment, policy enforcement (for bias, fairness, content safety), documentation, compliance with domain-specific regulations (e.g., medical privacy, financial regulations), user-consent mechanisms.

○ Domain-Adaptation and Customization Layer: mechanisms for domain-specific fine-tuning, language adaptation (e.g., multilingual support), embedding domain heuristics or constraints (e.g., medical ontologies, financial rules), integration with external risk models or knowledge bases, domain-specific prompt templates and guardrails.

5. Each layer interacts with others, forming a robust architecture for deploying LLMs across domains while ensuring performance, reliability, compliance, and adaptability.

6. Illustrative Deployment Workflow

We outline an example workflow using this framework:

○ Developers/researchers commit code or model-configuration changes to version control.

○ The CI/CD pipeline triggers automated training or fine-tuning, using the infrastructure layer (cloud GPUs) and retrieving versioned datasets and base model checkpoints from the registry.

○ Upon completion, automated evaluation suite in the evaluation layer runs domain-relevant benchmarks: e.g., for a medical LLM, runs clinical understanding tests; for finance LLM, runs financial scenario tasks.

○ If evaluation passes thresholds (for accuracy,

bias/fairness, performance), the pipeline deploys the model to a staging environment, where performance monitoring and logging capture runtime metrics.

○ Governance workflows — such as approval gates and audit logging — ensure compliance before promotion to production.

○ In production, monitoring and logging continue, with drift detection, feedback loop integration (user feedback, error reporting), and scheduled re-training or fine-tuning as necessary (e.g., to adapt to new data, drift, or domain changes).

7. Benefits and Challenges

Benefits:

○ Reproducibility and traceability: versioning of model, data, and code ensures experiments and deployments are reproducible.

Scalability:

cloud-based infrastructure and automated pipelines allow scaling to large user bases and high throughput.

○ Consistency and reliability: automated evaluation and testing reduce human error, ensure adherence to standards, and facilitate rapid updates.

○ Domain-specific compliance: governance and domain-adaptation layers ensure regulatory, ethical, and performance requirements are met.

○ Maintainability: modular architecture allows swapping or upgrading components (e.g., integrating new evaluation benchmarks, adding support for new domains).

8. Challenges and Risks:

○ Resource demand and cost: cloud GPU/TPU usage, storage, and compute costs can be high, especially for large models.

○ Complexity: building and maintaining such a layered framework demands skilled personnel and proper engineering practices; may be challenging for smaller organizations.

○ Domain evaluation limitations: domain-specific benchmarks may not exist, or may be limited in scope, raising risk of overlooked biases or failures.

○ Data privacy, security, and compliance: storing, logging, or fine-tuning on sensitive data (e.g., medical or financial) requires robust safeguards, which may complicate design.

○ Latency, performance tradeoffs: ensuring low latency for real-time applications may conflict with heavy logging, monitoring, or safety filters.

○ Model drift, degradation: continual learning or re-training may lead to unintended behavior, overfitting, or loss of previously acceptable performance.

Discussion

The proposed integrative framework sits at the confluence of two rapidly evolving areas: the widespread deployment of large language models, and the emergence of MLOps and CI/CD practices tailored to machine learning. By synthesizing these literatures and embedding domain-specific considerations, this framework aims to address the real-world challenges of operationalizing LLMs in production.

One of the core contributions is providing a modular, layered architecture that is domain-agnostic yet sufficiently flexible to accommodate domain-specific adaptation and evaluation. This is important because LLM use cases vary drastically across domains — the requirements, constraints, evaluation criteria, and regulatory landscape differ significantly. For instance, deploying an LLM-based assistant in clinical environments demands rigorous safety, fairness, and compliance checks (Gao et al., 2023; Dada et al., 2024), while financial applications require interpretability, auditability, risk assessment, and compliance with financial regulations (Lee et al., 2024). Multilingual applications demand fairness across languages and cultures (Yuan et al., 2023), and multimodal models — such as vision-language LLMs — bring in entirely different evaluation dimensions (Wang et al., 2024). Therefore, a one-size-fits-all deployment pipeline grounded only in generic metrics (e.g., perplexity) is insufficient.

The integration of MLOps practices and CI/CD processes into this framework brings several advantages. First, it ensures reproducibility and traceability: every model, dataset, and training configuration is versioned and tracked. Second, it supports rapid iteration and continuous improvement: as models evolve, bug fixes, performance enhancements, or domain adaptations can be built, tested, and deployed in a controlled manner. Third, it lends professionalism and maturity to model deployment: engineering standards, automated testing,

compliance, and governance are vital for real-world systems. These characteristics are increasingly important as organizations scale LLM systems, integrate them into business workflows, and face legal, ethical, or liability risks.

Nevertheless, this conceptual framework comes with limitations and caveats. Because it is based entirely on literature synthesis, without empirical deployment in a real-world setting, its practical effectiveness remains to be validated. Real-world constraints — such as cloud cost limits, latency requirements, organizational inertia, lack of domain-specific benchmark datasets, data privacy regulations — may hinder its implementation. Additionally, domain-specific evaluation benchmarks may simply not exist for some use cases, especially novel or niche applications; creating robust evaluations may be as difficult as building the models themselves.

Another concern relates to model drift and continual learning: while automated pipelines and re-training can help keep models up to date, they may also introduce unintended behavior, risk of catastrophic forgetting, or degradation in previously acceptable performance. Furthermore, privacy and compliance must be carefully managed: logging user inputs and outputs, or fine-tuning on user data (especially in domains like medicine or finance), can raise data-protection issues, regulatory liability, and ethical concerns. Governance mechanisms need to be robust, transparent, and enforceable; but designing such mechanisms is complex, particularly in organizations without mature ML governance practices.

Looking to the future, several promising directions emerge. First, empirical validation: deploying pilot systems based on this framework in different domains (medical, financial, multilingual, multimodal) can yield valuable insights, uncover pitfalls, and refine the framework. Second, development of domain-specific benchmark suites: for medicine (clinical safety, ethics), finance (risk, compliance), multilingual fairness, multimodal robustness, etc. Third, research into privacy-preserving MLOps practices: differential privacy, secure multiparty computation, federated learning pipelines integrated into the CI/CD workflow. Fourth, integration of human feedback loops, human-in-the-loop governance, and continual monitoring to catch emergent issues, bias drift, or performance degradation.

Finally, as LLM architectures evolve — with more efficient, smaller, or specialized models; or hybrid language-knowledge networks; or multimodal transformer variants — the framework must evolve accordingly. But because it is modular and layered, it is well positioned to adapt: new model types or evaluation metrics can be plugged in; new domain modules can be added without re-engineering the entire system.

In sum, we believe the proposed framework provides a timely, flexible, and robust blueprint for organizations seeking to deploy LLMs at scale and responsibly.

Conclusion

The remarkable rise of large language models has opened up unprecedented opportunities across fields such as natural language processing, medicine, finance, multilingual translation, and multimodal AI. Yet realizing this potential in production settings demands more than powerful models: it requires mature, scalable, reproducible, and governed deployment processes — an area where traditional machine learning practices often fall short. Through comprehensive literature synthesis, this article identifies a critical gap: the absence of an integrated, domain-agnostic but customizable framework for LLM deployment grounded in MLOps and CI/CD methodologies.

We propose a modular, layered framework that unifies infrastructural, operational, evaluation, governance, and domain-adaptation components, offering a blueprint for sustainable, scalable, and compliant deployment of LLMs across domains. The framework supports versioning, continuous delivery, domain-aware evaluation, monitoring, and governance, while remaining adaptable to domain-specific constraints and requirements. Although conceptual, the framework's design draws directly from peer-reviewed publications, surveys, and industry reports, ensuring it is rooted in current best practices and challenges described by researchers and practitioners.

We hope this framework serves as a foundation for empirical deployment, further research, and community-driven refinements. As LLMs continue to evolve, and as organizations increasingly integrate them into critical systems, robust, structured, and ethical deployment practices — as outlined in this framework — become not just beneficial, but essential.

References

1. Chandra, R. (2025). OPTIMIZING LLM PERFORMANCE THROUGH CI/CD PIPELINES IN CLOUD-BASED ENVIRONMENTS. International Journal of Applied Mathematics, 38(2s), 183-204.

2. John Snow Labs. (2024). Introduction to Large Language Models (LLMs): An Overview of BERT, GPT, and Other Popular Models. Available online: https://www.johnsnowlabs.com/introduction-to-large-language-models-llms-an-overview-of-bert-gpt-and-other-popular-models/ (accessed on 14 September 2024).
3. Gao, Y.; Baptista-Hon, D. T.; Zhang, K. (2023). The inevitable transformation of medicine and research by large language models: The possibilities and pitfalls. MEDCOMM-Future Med, 2, 1–2.
4. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. (2024). Large language models: A survey. arXiv:2402.06196.
5. Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Chen, H.; Dou, Z.; Wen, J. R. (2023). Large language models for information retrieval: A survey. arXiv:2308.07107.
6. Lee, J.; Stevens, N.; Han, S. C.; Song, M. (2024). A survey of large language models in finance (FinLLMs). arXiv:2402.02315.
7. Yuan, F.; Yuan, S.; Wu, Z.; Li, L. (2023). How Multilingual is Multilingual LLM? arXiv:2311.09071.
8. Dada, A.; Bauer, M.; Contreras, A. B.; Koraş, O. A.; Seibold, C. M.; Smith, K. E.; Kleesiek, J. (2024). CLUE: A Clinical Language Understanding Evaluation for LLMs. arXiv:2404.04067.
9. Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. (2024). VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. In Proceedings of the 37th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 6–10 December 2024.
10. Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. (2023). A comprehensive overview of large language models. arXiv:2307.06435.
11. Pahune, S.; Chandrasekharan, M. (2023). Several categories of large language models (LLMs): A short survey. arXiv:2307.10188.
12. Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; … Wen, J. (2023). A survey of large language models. arXiv.
13. Berberi, L.; Kozlov, V.; Nguyen, G.; Díaz, J. S.; Calatrava, A.; Moltó, G.; … García, Á. L. (2025). Machine learning operations landscape: platforms and tools. Artificial Intelligence Review, 58(6).
14. Zarour, M.; Alzabut, H.; Alsarayrah, K. (2025). MLOps best practices, challenges and maturity models: A systematic literature review. Information and Software Technology, 107733.
15. Kazmierczak, J.; Salama, K.; Huerta, V. (2024, August 28). MLOps: Continuous delivery and automation pipelines in machine learning. Google Cloud.