Towards a Standardized Business Process Model for LLMOps

Maria Chernigovskaya[1][a], Damanpreet Singh Walia[2][b], Ksenia Neumann[2][c], André Hardt[1], Abdulrahman Nahhas[1][d] and Klaus Turowski[1]

[1]*MRCC VLBA, Otto-von-Guericke University, Universitaetsplatz 2, Magdeburg, Germany*

[2]*BIRD Lab, Otto-von-Guericke University, Universitaetsplatz 2, Magdeburg, Germany*

{*maria.chernigovskaya, damanpreet.walia, ksenia.neumann, klaus.turowski, andre.hardt, abdulrahman.nahhas*}*@ovgu.de*

Keywords: LLM, Large Language Model Operations, LLMOps, Standardization, Process Model, BPMN.

Abstract: The generalization and standardization of the Large Language Model Operations (LLMOps) life cycle is crucial for the effective adoption and management of Large Language Models (LLMs) in a business context. Researchers and practitioners propose various LLMOps processes however they all tend to lack formalization in their design. In this paper, we address the absence of a standard LLMOps model for enterprises and propose a generalized approach to adopting LLMOps into existing enterprise system landscapes. We start by identifying the state-of-the-art LLMOps processes through a systematic literature review of peer-reviewed research literature and gray literature. Considering the scarcity of relevant publications and research in the area discovered during the initial stage of the research, we propose a generic, use-case-agnostic, and tool-agnostic LLMOps business process model. The proposed model is designed using the Business Process Model and Notation (BPMN) and aims to contribute to the effective adoption of LLM-powered applications in the industry. To the best of our knowledge, this paper is the first attempt to systematically address the identified research gap. The presented methods and proposed model constitute the initial stage of the research on the topic and should be regarded as a starting point toward the standardization of the LLMOps process.

1 INTRODUCTION

Large Language Models (LLMs) have emerged as a technological breakthrough in the field of Artificial Intelligence (AI). Advanced neural network models trained on vast amounts of data have demonstrated human-like performance in complex natural language processing tasks like sentiment analysis (Chen, 2024), text generation (Liang et al., 2024), and translation (He et al., 2023). Due to its remarkable capabilities, enterprises began looking for ways to integrate this powerful tool into their IT landscape and existing business processes. To achieve effective and seamless integration of LLMs, Large Language Model Operations (LLMOps) must be applied. LLMOps is a set of engineering practices that emerged as an extension of the Machine Learning Operations (MLOps) paradigm to regulate the lifespan of LLM-powered applications (Diaz-De-Arcaya et al., 2024). Standardization of LLMOps may provide substantial advantages to enterprises seeking to efficiently deploy and manage LLMs by ensuring homogeneity in deployment and maintenance across various teams and projects.

The standardization and formalization of the LLMOps framework could be achieved conceptually and later practically through various standard modeling languages (e.g., Unified Modelling Language (UML), Systems Modelling Language (SysML), etc.). However, to capture not only the technology part but also its integration with the ongoing business processes, the use of the Business Process Model and Notation (BPMN) becomes sensible, as it has been the prevailing business process modeling language standard since its debut in 2004 (Respício and Domingos, 2015). Additionally, adopting BPMN practices could contribute significantly to the standardization of LLMOps by supplying a systematic framework for deploying, fine-tuning and monitoring every step involved in the LLM-powered application life cycle.

In this paper, we propose an use-case-agnostic and tool-agnostic BPMN model to standardize the process of adopting LLMOps into existing IT landscape. To the best of our knowledge, this work is the first attempt to address this problem systematically, while

[a] https://orcid.org/0009-0004-2763-6970
[b] https://orcid.org//0009-0002-4044-5613
[c] https://orcid.org/0000-0002-3713-8893
[d] https://orcid.org/0000-0002-1019-3569

Chernigovskaya, M., Walia, D. S., Neumann, K., Hardt, A., Nahhas, A. and Turowski, K.
Towards a Standardized Business Process Model for LLMOps.
DOI: 10.5220/0013377700003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 856-866
ISBN: 978-989-758-749-8; ISSN: 2184-4992

highlighting the state-of-the-art research findings. In our work, we aim to answer the following Research Questions (RQs):

RQ 1. What are the state-of-the-art LLMOps processes proposed by academics and industry practitioners?

RQ 2. How can a standardized LLMOps process model be developed using identified state-of-the-art processes and BPMN?

The main contribution of this paper is an easy-to-adopt BPMN LLMOps process model for business use. The resulting process model is an artifact, based on the definition of artifact by (Hevner et al., 2004). We derive use-case- and tool-agnostic processes from the existing LLMOps processes being employed by researchers and practitioners to ensure a compatible process model to establish the foundation for subsequent works on standardizing the LLMOps life cycle. Overall, we assert two claims: First, we demonstrate the lack of a standard LLMOps process model through a critical viewpoint on distilled information from a literature review. Second, we propose LLMOps standardized process model that is generic and ensures compatibility for easy adoption.

The paper proceeds as follows: in the subsequent section 2 we systematically explore the state-of-the-art LLMOps processes in the peer-reviewed and gray literature with a goal to answer RQ 1.. The section also provides a critical viewpoint on the shortlisted literature, where the main limitations of the analyzed works are discussed. In section 3 we present a consolidation of the identified in the literature LLMOps and propose a design of a standardized LLMOps process model, addressing RQ 2.. In the section 4 we once again discuss the goal of the proposed work and address its main limitations. We summarize the main findings and contribution of this work and discuss future research perspective in the section 5.

2 STATE-OF-THE-ART LLMOps

To the best of our knowledge, no study has been conducted aiming to achieve the same research goal as ours i.e. to propose a standardized process model of LLMOps. To conduct the research systematically, the method of Systematic Literature Review (SLR) has been chosen, as suggested by (Okoli, 2015) and (Kitchenham et al., 2004), as it provides a structured approach of the relevant literature analysis and helps to summarize its findings. Thus, a comprehensive SLR has been performed in order to derive a LLMOps standardized process model based on existing processes and models proposed in the related research.

2.1 Literature Search

This section describes in a detail process of conducting an SLR based on the guide by (Okoli, 2015) on the highlighted topic as well as its findings. The visual representation of the conducted SLR is depicted in Figure 1 and it consists of five screening stages: (0) retrieving articles based on a constructed out of keywords search string from various databases, (1) applying exclusion criteria, (2) reading title and abstract, (3) reading full-text, (4) synthesizing peer-reviewed sources with gray literature such as white papers. Each stage consists of one or more steps and displays remaining records in each stage. These steps follow the subsequent execution order: selecting literature sources (scientific databases), defining a search string, retrieving articles, removing duplicates, several screening stages of the articles according to the defined exclusion and inclusion criteria, enriching the intermediate results by extending our SLR through forward and backward search and gray literature.

The first step of conducting an SLR is to define a search string that covers the first research question. The constructed query is based on the identified keywords and formulated as following: ((”LLM*” OR ”Large language model*”) AND (”process model*” OR ”Operation*”)) OR (”LLMOps” OR ”LLM-OPs” OR ”Large language model* operation*”). A broad set of databases is chosen in order to ensure a high level of inclusion. Query search was performed on the 9th of October, 2024 over thirteen databases: ACM, Springer Link, Scopus, IEEE Xplore, ScienceDirect, Web of Science, AISel, Scitepress, mdpi, Wiley, Taylor&Francis, Emerald Insight and Sage, accumulating to 43 initially retrieved articles.

Our exclusion criteria: complete duplicates, not peer-reviewed, closed access book, language other than English, title / abstract / full-text not related as per PRISMA list (Tricco et al., 2018). The complete overview of the search process including yielded results is depicted by Figure 1. Firstly, out of the 43 results eight duplicates were removed. Out of the remaining 35 four were removed as they were not peer-reviewed and eight due to original source being in closed access. Followed by the filtration stages of title and abstract relevance, 17 records were removed, resulting into six full-text-related results. As the final number of records happened to be too small, we decided to extend the peer-reviewed literature through forward and backward search (two additional sources found), white papers (accumulating five), three books,

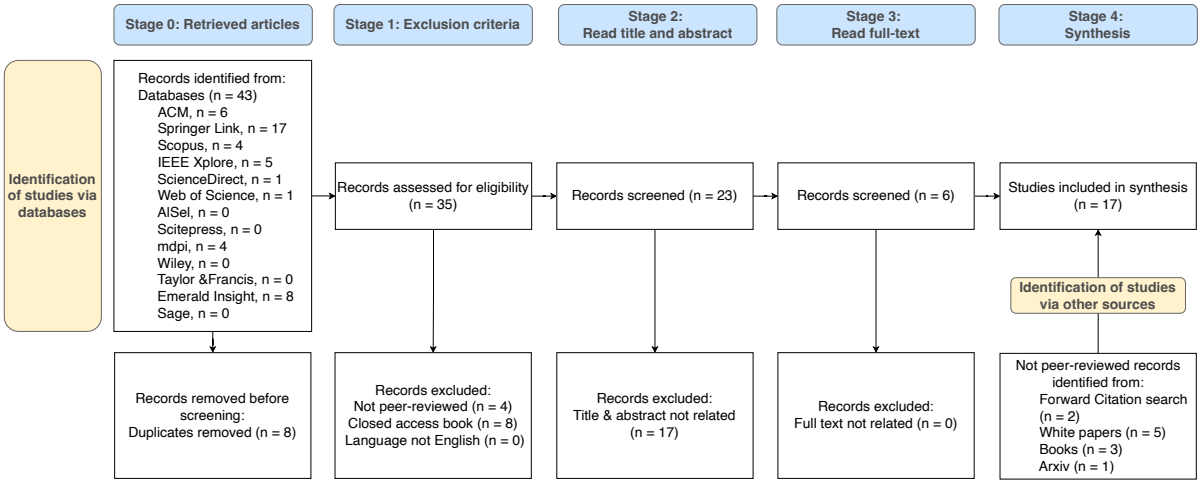

Figure 1: Synthesis of the conducted SLR process with the corresponding screening stages (Stage 0 - Stage 4).

and a single article from Arxiv. White papers were chosen based on the market share of the publishing company selecting following prominent companies such as AWS, Amazon Web Services, Fujitsu, AMD epyc, Dell Technologies, Nvidia, appliedAI, salesforce, RedHat and Intel. However, after the conducted SLR, the final set of relevant white papers was reduced to five (Fujitsu, 2024), (Basak, 2024), (Kartakis and Hotz, 2023), (Datta et al., 2024), (Venkatapathy, 2023). Some leading LLM vendors (e.g., Open AI[1], Google Gemini[2], etc.) were also considered in the initial stages of the SLR. However, these companies were not included in the final set as they tend to focus more on the products' performance rather than on their integration into existing business processes.

2.2 Literature Synthesis

In order to combine our findings, a thematic synthesis as proposed by (Cruzes and Dyba, 2011) was applied. Their multi-step approach helps to identify recurring themes from multiple studies, analyze and interpret them, so the conclusions can be drawn in the systematic reviews. In other words, it synthesizes the findings identified within primary studies. After the SLR was completed and full-text analysis of the final results was conducted, four main groups aka "themes" were identified within the final set of studies. The proposed categorization consists of shortlisted literature from Stage 4 of SLR and is based on RQ 1. to identify state-of-the-art LLMOps processes with additional categories for LLMOps tool(s), LLMOps utilization and LLMOps related actors. The overview of all four groups can be found in Table 1.

The first group consists of thirteen studies and describes either LLMOps processes or their stages. The second group is assigned to category "LLMOps tool(s)" and contains studies, in which authors presented or discussed at least one LLMOps tool. In the third identified category, all studies that describe LLMOps utilization, were grouped. Finally, the fourth category, where the authors described LLMOps actors and their responsibilities, concludes Table 1 with just two studies. It it worth mentioning, that some of the publications fall under multiple thematic categories as they cover multiple topics at the same time.

2.3 Critical Viewpoint

To evaluate the quality of the conducted review, we present our perspectives on the literature within the "LLMOps processes" cluster. From the perspective of AWS, one of the leading cloud service providers, (Kartakis and Hotz, 2023) highlights the distinction between MLOps, Foundational Model Operations, and LLMOps. The authors also detail the processes, actors, and tools involved in LLMOps, providing practitioners with a comprehensive framework that encompasses all critical dimensions relevant to businesses. Expanding on this work, (Basak, 2024) presents a guide for the operation and management of LLMs using Apache Airflow. While both works effectively delineate the various processes, roles, and responsibilities associated with LLMOps, they do not provide detailed work and adoption analysis is also missing. Furthermore, the literature demonstrates a noticeable preference for tools within AWS's service catalog, such as Amazon Managed Workflows for Apache Airflow.

The white paper by TruEra and Intel (Datta et al.,

[1]https://openai.com/[03.02.2025]

[2]https://gemini.google.com/[Accessed on 03.02.2025]

Table 1: Categorization of the identified studies based on the thematic category.

Category	Literature
LLMOps processes	(Basak, 2024), (Datta et al., 2024), (Diaz-De-Arcaya et al., 2024), (Huang et al., 2024), (Kartakis and Hotz, 2023), (Kamath et al., 2024), (Park et al., 2024), (Parnin et al., 2023), (Reddy et al., 2024), (Roychowdhury, 2024), (Shan and Shan, 2024), (Venkatapathy, 2023)
LLMOps tool(s)	(Arawjo et al., 2024), (Basak, 2024), (Datta et al., 2024), (Fujitsu, 2024), (Huang et al., 2024), (Kamath et al., 2024), (Kartakis and Hotz, 2023), (Park et al., 2024), (Venkatapathy, 2023), (Wang and Zhao, 2024)
LLMOps utilization	(Chen, 2024), (Kamath et al., 2024), (Parnin et al., 2023), (Shan and Shan, 2024), (Shi et al., 2024), (Venkatapathy, 2023)
LLMOps actors and responsibilities	(Basak, 2024), (Kartakis and Hotz, 2023)

2024) provides a detailed exploration of the LLMOps workflow, consolidated based on the emerging LLMOps technology stack available in the market. The authors acknowledge the limited listed technology stack while emphasizing that the workflow's development, grounded in these emerging tools, enhances its validity and achieves tool-agnosticism by encompassing a variety of technologies. Furthermore, the white paper's broad focus on LLM use cases adds to its generalization and potential for wider adoption. However, the workflow presented is predominantly technical in nature, highlighting a notable absence of generalized approach and relevant strategic processes.

The authors of (Diaz-De-Arcaya et al., 2024) consolidate the definition of LLMOps and its various stages through a synthesis of existing literature. Their unified definition emphasizes the need for LLMOps, presenting it as a customized MLOps approach tailored to address business challenges such as cost management, technical hurdles, and the selection of tools and infrastructure. The authors also identify and comprehensively define the stages of LLMOps. However, the absence of a visual model and detailed information on the sequence of stages is a notable limitation. While the stage numbering implies a sequential flow, this is not explicitly clarified. Furthermore, the correlation between the defined stages and the strategic challenges faced by businesses mentioned in the LLMOps definition appears incomplete, as these challenges are not adequately reflected in the stages. Although the model presented is sufficiently general for LLM use cases, the lack of detailed information on tools makes it challenging to assess its tool-agnostic nature.

A shortlisted book, (Huang et al., 2024), provides an explanation of LLMOps, emphasizing its needs and benefits. It effectively highlights the differences between MLOps and LLMOps. Regarding LLMOps processes, the book focuses on and elaborates upon processes related to LLM security. Additionally, it lacks a visual representation of a process model. Another shortlisted book, (Kamath et al., 2024), discusses certain aspects of LLMOps and presents an example of an LLMOps architecture for an LLM-based chatbot. However, the proposed architecture is complex and overly detailed, and its lack of compliance with standard modeling notations makes it challenging to interpret at a strategic organizational level.

(Park et al., 2024) introduces an LLMOps pipeline, termed "*LlamaDuo*", designed for migration from service-oriented LLMs to smaller, locally manageable models. While the proposed pipeline shares some similarities with other shortlisted LLMOps processes, its objectives differ from the intended focus of this work. In (Parnin et al., 2023), the authors gathered challenges and workflows for developing copilot-like products through interviews with professional software engineers. The iterative process model outlined in the study describes all underlying processes by incorporating the perspectives of these practitioners. Although the process model does not employ any standard modeling notation, it effectively provides a comprehensive representation of LLMOps processes tailored to this specific use case.

The causes and impacts of hallucinations in LLMs are examined in (Reddy et al., 2024), along with strategies to mitigate this issue. While the article also presents an LLMOps process model, it lacks a clear explanation of its origin and underlying rationale. This drawback limits its applicability and leads to the exclusion of this process model from further consideration in this study. The short article (Roychowdhury, 2024) proposes a three-stage LLMOps process model for finance-focused LLM products, starting with the definition of the business case. However, as the processes are presented in an abstract manner and lack

detailed explanations, it is challenging to achieve a comprehensive understanding of the model.

The research article (Shan and Shan, 2024) introduces the 4D LLMOps process model, outlining best practices and application scenarios for implementing LLMOps in enterprises. The authors also provide a list of potential tool stacks that can be utilized within the proposed framework. However, the rationale behind the naming of the stages is unclear, as the terms do not maintain a consistent level of abstraction. For instance, "Deploy" is a fundamental operational task, while "Deliver" encompasses multiple operational activities. Furthermore, the model does not adhere to standard modeling notations, which may limit its clarity and adoption. The white paper by Dell Technologies (Venkatapathy, 2023) provides a comprehensive outline of the validated life cycle design for generative AI in retail use cases. The author thoroughly explains all elements involved in the design and presents a tech stack based on NVIDIA technologies. However, the life cycle lacks adherence to standard modeling notations, which hinders its simplification and broader adoption.

In summary, the definition of LLMOps provided by (Diaz-De-Arcaya et al., 2024) establishes the criteria for assessing the completeness of LLMOps lifecycle models. Among the twelve reviewed sources on LLMOps processes, only seven—(Basak, 2024), (Datta et al., 2024), (Diaz-De-Arcaya et al., 2024), (Kartakis and Hotz, 2023), (Parnin et al., 2023), (Roychowdhury, 2024), and (Shan and Shan, 2024)—present well-defined LLMOps models that align with the focus of this research. The absence of a comprehensive LLMOps process model with standardized notation and consistent terminology remains evident. A detailed assessment of the identified models for standardization is provided in the following section.

3 LLMOPS STANDARDIZED PROCESS MODEL

To assess the level of standardization in LLMOps business process models, we performed a subjective evaluation of their completeness and generality due to the absence of a predefined set of criteria. For completeness, we evaluated whether the LLMOps models incorporate both strategic and technical aspects of the LLMOps life cycle, as defined in (Diaz-De-Arcaya et al., 2024), and whether they effectively address key business challenges. For generality, we examined three key factors:

- **Use of Standard Modeling Notation.** To ensure ease of interpretation and clarity.
- **Use-Case-Agnostic.** To facilitate adoption across a diverse range of LLM use cases.
- **Tool-Agnostic.** To support unbiased adoption by accommodating a wide variety of available LLMOps tools.

This approach allows us to systematically evaluate the strengths and limitations of the proposed models.

The presented comparison in Table 2 aims to assess generality of the identified state-of-the-art LLMOps models from the literature. Here, columns represent generality criterion: Use of Standard Notation, Use-Case-Agnostic and Tool-Agnostic. Literature citation is used as identifier for identified LLMOps models. "+" and "−" used as marking scheme showing compliance and non-compliance respectively. As it can be observed, none of the identified publications presents a fully standardized approach for generalizing LLMOps. Additionally, none of the reviewed literature utilized BPMN in their proposed artifacts. Regarding the completeness of LLMOps models, (Parnin et al., 2023) and (Shan and Shan, 2024) offer the most comprehensive process models compared to other identified approaches.

3.1 Redesigning LLMOps Life Cycle

The seven shortlisted LLMOps models were assessed against the aforementioned criteria for standardizing LLMOps process models. However, none of the identified models fully met the criteria. Consequently, a method was devised to redesign the LLMOps life cycle from the identified models, resulting in a standardized LLMOps process model.

Firstly, process models from the shortlisted literature were consolidated, as illustrated in Figure 2, to redefine processes into consistent process group categories (as shown in legend) to meet the requirement of RQ 2., generic LLMOps process model artifact. Here, the proposed process group's color is used to mark processes of models based on the responsible activity areas. This classification aids businesses in identifying relevant roles for these activities.

Secondly, six process groups were identified within the shortlisted models, providing a high-level overview of the LLMOps life cycle:

1. **Strategic.** Processes focused on defining and formulating business needs.
2. **Data.** Processes related to data collection, processing, transformation, and management.
3. **LLM.** Processes centered on the selection, fine-tuning, and evaluation of LLMs.

Table 2: Comparison table to assess the generality of the identified state-of-the-art LLMOps models from the literature.

Literature	Generic LLMOps Model		
	Use of Standard Notation	Use-Case-Agnostic	Tool-Agnostic
(Basak, 2024)	–	+	–
(Datta et al., 2024)	–	+	+
(Diaz-De-Arcaya et al., 2024)	–	+	–
(Kartakis and Hotz, 2023)	–	+	–
(Parnin et al., 2023)	–	–	–
(Roychowdhury, 2024)	–	–	–
(Shan and Shan, 2024)	–	+	+

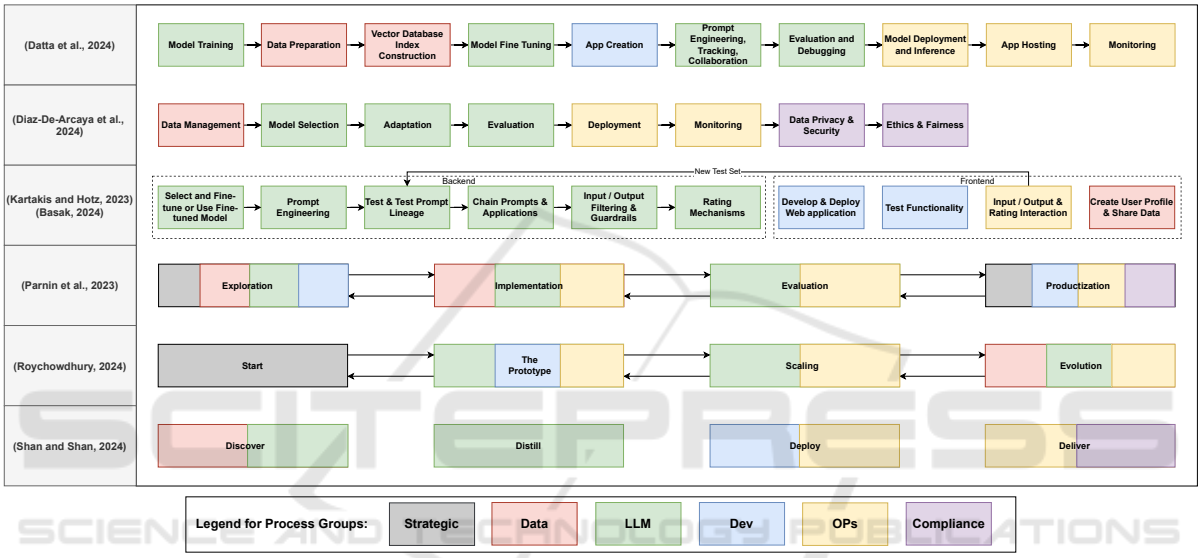

Figure 2: Consolidated view of the identified LLMOps models with colored areas of responsibilities.

4. **Development (Dev).** Processes involving the integration of LLMs into system infrastructures.
5. **Operations (Ops).** Processes responsible for delivering and maintaining the product to end users.
6. **Compliance.** Processes ensuring periodic audits and adherence to ethics, fairness, data privacy, and security.

Thirdly, these process groups and their underlying processes are abstractly represented in Figure 3, which depicts an infinite knot diagram of LLMOps process model as per aforementioned redefined process groups on an abstract level. Color consistency is maintained with Figure 2. This diagram presents the sequence of processes in the LLMOps life cycle, drawing inspiration from the MLOps model outlined in the following ML4Devs article[3]. Lastly in the subsequent subsection, this LLMOps process model is further formalized and standardized using BPMN.

[3]https://www.ml4devs.com/articles/mlops-machine-learning-life-cycle/ [Accessed on 14.12.2024]

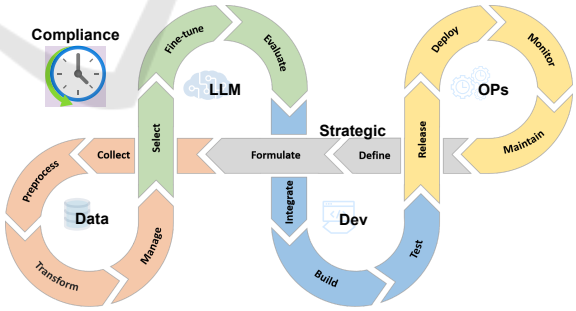

Figure 3: Infinite loop representation of LLMOps process model derived from Figure 2.

3.2 LLMOps BPMN Model

The BPMN model depicted in Figure 4 is the central artifact of this paper. It aims to generalize and standardize the steps required to effectively integrate LLMOps into various business scenarios. The model summarizes the findings of analyzed scientific and gray literature and reflects the consensus among

academia and industry with respect to the LLMOps workflow definition. Finally, the presented artifact aims to highlight the use-case and tool-agnostic nature of the proposed LLMOps representation.

The presented model consists of one main pool and a black box pool that depicts an IT infrastructure. The main pool of the model describes the complete LLMOps workflow within the organization divided into two lanes, each illustrating a distinct functionality level: strategic and operational. Each defined level is represented by a team that performs the respective roles. The strategic team consists of key decision-makers in charge of defining goals and objectives, developing strategies, and supervising the overall execution. The operational team is responsible for carrying out day-to-day tasks that align with the business strategy. In this scenario, we consider a development team to be a part of the operational team as it is responsible for performing tasks that are operational in nature. Depending on the size and type of the enterprise, the exact roles and activities of both teams may differ.

The decision to illustrate the IT infrastructure as a black box entity is motivated by two main reasons. Firstly, viewing IT infrastructure as a black box enables our model to stay general enough and adaptable to various infrastructure configurations (on-premises, cloud, hybrid, etc.). Secondly, the goal of the model is to outline key touch points between two pools and the associated impact of the infrastructure without delving into its internal activities and technicalities as both will strongly depend on the specific technology stack and their vendors.

The process starts with the launch of the LLMOps project and an outline of its general scope and objectives. Here, the desired outcome of the project, that depends on the application of an LLM, must be clearly defined. It could, for example, be targeting the enhancement of the customer experience, automation of tasks, improvement of decision-making, etc. Defined goals are subsequently translated into business requirements, which typically comprise objectives, benefits, constraints, value propositions, stakeholder expectations, and use case descriptions. The defined business requirements are then passed onto the operational team to derive functional requirements. Functional requirements are technical specifications that engineers and developers must follow to implement the expected functionalities, while aligning with business and non-functional requirements. Requirement engineering is an important activity, especially with regard to complex systems, as it guarantees that all parties involved have a shared understanding of the project objectives and expectations.

Once functional and non-functional requirements are defined, the operational team proceeds with the design of artifact architecture. The architecture design is a foundational step, that lays out a blueprint of how exactly the application will be integrated into the existing system. It outlines the application's high-level structure and defines its main components, their functions, and connections with one another. The proposed architecture is passed onto the strategical team to evaluate its alignment with the defined objectives. If the proposed architecture is approved, the blueprint is finalized and passed on to development activities. If the provided architecture does not pass the approval step, it is subjected to further refinements until it satisfies the set requirements.

In alignment with the strategic workflow, the operational team proceeds with the activities focusing on the data management and preparation. At this step, the relevant data is collected, cleaned, normalized, and transformed to the form suitable for usage within an LLM model. Additionally, the data quality evaluation process is performed. Data preparation step is critical as it ensures quality, validity, and usefulness of the used data throughout the life cycle. Following this step, the operational team selects a suitable pre-trained LLM model that best meets the functional and non-functional requirements and is compatible with the business scenario. At this step, factors like model size, model type, costs (e.g., personnel, licensing, infrastructure), and technical requirements must be considered. Within this work we assume that a typical enterprise, due to the lack of expertise, high cost and significant computational complexity, would rather not invest in training their own LLM model and rather use a pre-trained one.

After selecting the base model, the decision to fine-tune LLM is made. A fine-tuning phase is required to tailor the model for a particular type of task and data. As shown in the proposed BPMN model, the fine-tuning stage can be skipped if the base model already performs sufficiently to serve the intended use. At this point, the strategic team activities (e.g., finalization of the project blueprints) integrate with the operational team's workflow, with a touch point right before the artifact development activity. Artifact development is a transformational phase that merges the approved architecture design and requirements to produce a functional LLM artifact integrated with all its components. Such components might include various APIs, databases, user interfaces, embedding tools, etc.

The subsequent evaluation stage examines whether the produced artifact fulfills the business, functional, and non-functional requirements before proceeding to the deployment phase. The assessment

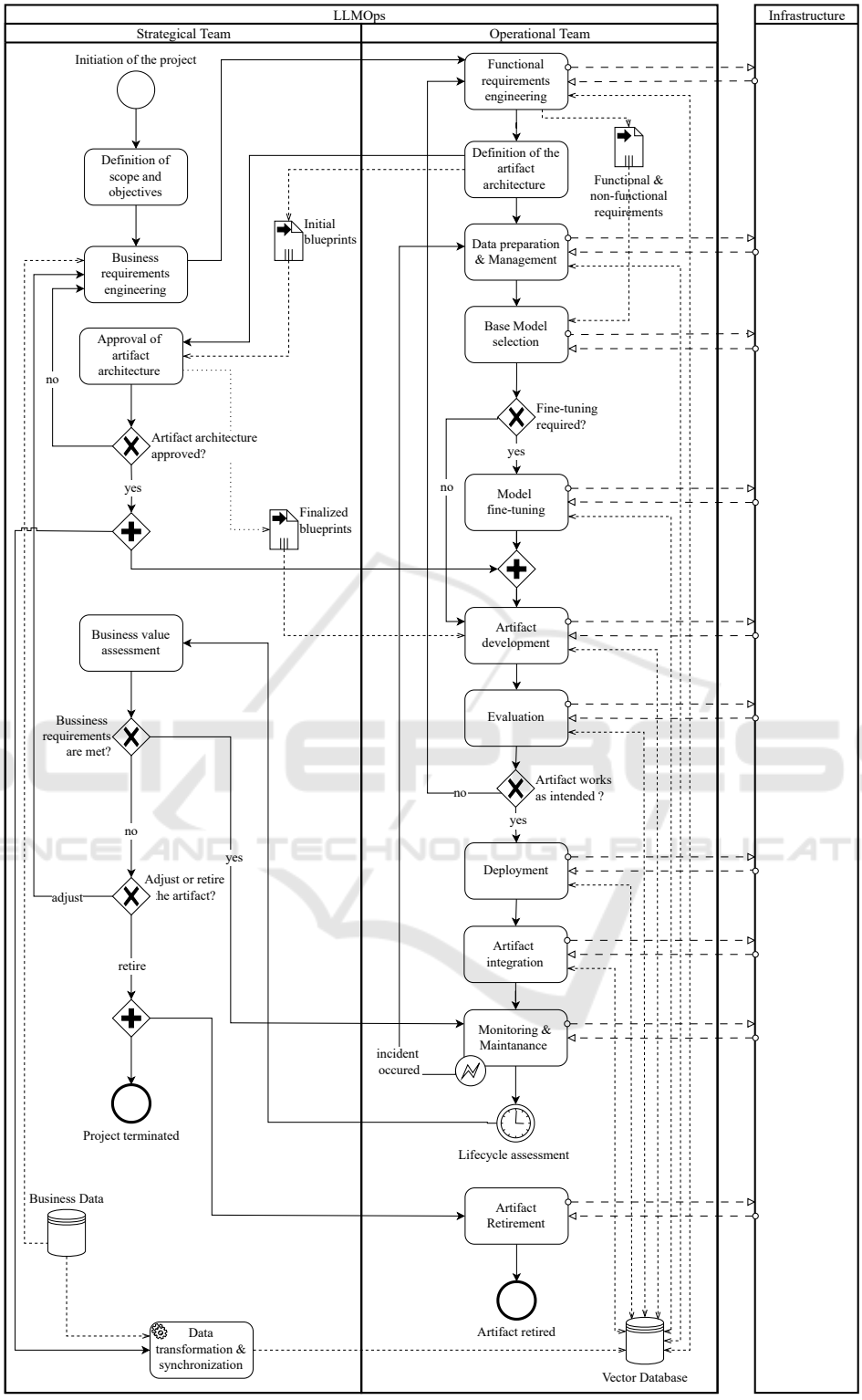

Figure 4: Standardized and generalized design of LLMOps process model using BPMN 2.0.

process might involve artifact performance testing (e.g., response time and throughput evaluation), output accuracy, etc. The evaluation process consists of several cycles of the artifact assessments and culminates in two possible outcomes: a verified artifact suitable for deployment or additional refining if inadequacies or inconsistencies with the requirements are discovered. If the artifact fails to perform as intended, the root cause has to be identified in the preceding activities, with the worst-case scenario necessitating the functional requirements to be refined.

After being evaluated, the artifact is deployed into the company's infrastructure. This step ensures that the artifact is technically functional and effectively integrates with running other applications, databases, and daily workflows. It usually involves setting up the software environment, allocating the required computing resources, and carrying out limited live tests to observe its behavior in real-world settings. The follow-up artifact integration activity is responsible for seamlessly integrating the artifact with the enterprise's current IT systems and processes. It includes integration with standard enterprise IT systems (e.g., Enterprise resource planning (ERP), Customer Relationship Management (CRM)), data synchronization between databases used for day-to-day business activities and the databases containing representations suitable for an LLM (e.g., vector storage), general workflow alignment, user interface integration, and data privacy compliance validation.

The following phase is the monitoring and maintenance of the deployed and integrated into the production environment artifact. This stage consists of two distinct activities with different sets of responsibilities. Monitoring involves tracking the artifact's activity in real-time to ensure the alignment of the artifact's behavior with the requirements defined in the earlier stages of the project. It comprises tasks such as performance monitoring, output quality assessment, anomaly and error detection, log analysis, and ensures continuous compliance with the data security and privacy regulations. Additionally, continuous maintenance of the software (e.g., security updates) and infrastructure components ensure that the artifact stays functional, manages incidents, and adheres to continuously evolving data and production environment realities (e.g., hardware generation changes and licensing changes). This phase includes activities such as regular artifact updates and correcting any potential errors detected during monitoring.

Once the artifact is integrated, its life cycle must be evaluated on a regular basis to ensure that it continues to provide the intended value while being relevant to evolving business requirements, significant technology developments (e.g., new model types, efficiency improvements), and user feedback. The life cycle assessment is additionally supported by a business value assessment, which serves as a systematic method applied to establish the project's continuous viability. This approach assists the strategic team in decision making on the project's extension, adjustments, or termination depending on the outcomes of the assessment.

If the artifact proves its practicality, it continues to operate until the next assessment cycle, or it might be determined that it requires specific adjustments to meet business requirements. If, based on the evaluation performed, the artifact is considered outdated or no longer valuable, it is retired accordingly, with all its dependencies being transitioned, archived, and finally decommissioned from the production environments. The exact decommissioning and archival procedures strongly depend on the nature of the project as well as the compliance regulations it falls under, if any. The project termination process is initiated concurrently with the artifact retirement.

4 DISCUSSION

Based on the key findings of the research, we can conclude that the relevant literature for the defined problem is notably sparse. Furthermore, none of the analyzed sources proposed a formalized or standardized depiction of LLMOps processes with the underlying activities. The majority of proposed representations in academic literature were predominantly broad and theoretical. Conversely, in industrial literature (white papers), the presented models were rather specific and tailored for the particular technological stacks. As a result, there was a lack of consistency and formalization in the LLMOps presentation in the collected literature.

To address the lack of consistency and generalization, we sought to consolidate the identified variations of LLMOps processes from academic and industrial fields into a single formalized representation. By analyzing the presented LLMOps processes, we selected stages of the process on which numerous authors appeared to have a consensus and consolidated them in Figure 2 and Figure 3. Following this approach, we built a research foundation for our business process model depicted in Figure 4.

The resulting business process model is the research artifact of this work that presents the main stages and activities required for LLMOps integration. We utilized the standard modeling notation like BPMN and formalized LLMOps workflow based on

the literature findings to achieve a high degree of generalization of the proposed model. By this approach we try to ensure its compatibility for easier adoption in the business contexts.

4.1 Limitations

Although the proposed artifact is based on an extensive and consistent literature analysis and designed using a standard modeling language, we acknowledge that we can not yet claim an overall standardization and generalization. To reach this point, a number of real-world use cases must be applied and tested to prove the proposed model's practical validity. We recognize this limitation and aim to address it in our future work.

5 CONCLUSION AND OUTLOOK

Successful integration of LLMs into an enterprise presents numerous challenges and necessitates the adoption of LLMOps. Standardizing LLMOps practices may provide significant advantages to businesses seeking to effectively manage the LLM life cycle while ensuring workflow consistency across various teams and projects. In this paper, we conducted an extensive SLR on the existing standard LLMOps models in both academic and industrial literature. Furthermore, we consolidated an overview of identified LLMOps processes and designed a cyclic LLMOps representation in Figure 2 and Figure 3 as a visual summary of the analyzed literature. Both figures are utilized as a trustworthy research foundation for our artifact.

Based on the SLR findings, we concluded that none of the discovered paper proposed a formalization or standardization model for LLMOps adoption. Moreover, none of the obtained findings relied on some standard modeling languages (e.g., BPMN, Unified Modeling Language). To address the identified research gap, we proposed an use-case and tool-agnostic LLMOps business model designed using BPMN. The model is designed to formalize and standardize the main steps required to effectively integrate LLMOps into various business scenarios. The conducted SLR and obtained findings answer RQ 1.. The proposed BPMN model in Figure 4 addresses and answers RQ 2..

As discussed previously, in the subsection 4.1, the presented work is subject to certain limitations that we intend to tackle in our future work. Firstly, we aim to test the proposed artifact on the real-world use cases from various domains to prove its generalization. The discovered findings might lead to further adjustments and refinements of the presented model. Secondly, we intend to elaborate more on specific type of activities and the related tasks and roles. Therefore, we believe that the presented results and artifact should not be regarded as final but rather considered as starting point.

REFERENCES

Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M., and Glassman, E. L. (2024). ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In Mueller, F. F., Kyburz, P., Williamson, J. R., Sas, C., Wilson, M. L., Dugas, P. T., and Shklovski, I., editors, *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM Digital Library, pages 1–18. Association for Computing Machinery.

Basak, P. (2024). Unlocking fmops/llmops with airflow: A guide to operationalizing and managing large language models. Technical report, Airflow Summit 2024.

Chen, T. (2024). Challenges and opportunities in integrating llms into continuous integration/continuous deployment (ci/cd) pipelines. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 364–367.

Cruzes, D. S. and Dyba, T. (2011). Recommended steps for thematic synthesis in software engineering. In *2011 international symposium on empirical software engineering and measurement*, pages 275–284. IEEE.

Datta, A., Sen, S., and Bandyopadhyay, A. (2024). Llmops explained. Technical report, TruEra and Intel Corporation.

Diaz-De-Arcaya, J., López-De-Armentia, J., Miñón, R., Ojanguren, I. L., and Torre-Bastida, A. I. (2024). Large language model operations (llmops): Definition, challenges, and lifecycle management. In *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–4. IEEE.

Fujitsu (2024). Fujitsu composite ai. Technical report, Fujitsu Limited.

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang, X. (2023). Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, pages 75–105.

Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S., and Ponnapalli, J. (2024). *Generative AI Security: Theories and Practices*. Springer.

Kamath, U., Keenan, K., Somers, G., and Sorenson, S. (2024). *Large Language Models: A Deep Dive*. Springer.

Kartakis, S. and Hotz, H. (2023). Fmops/llmops: Operationalise generative ai using mlops principles. Technical report, Amazon Web Services Inc.

Kitchenham, B. A., Dyba, T., and Jorgensen, M. (2004). Evidence-based software engineering. In *Proceedings. 26th International Conference on Software Engineering*, pages 273–281. IEEE.

Liang, X., Wang, H., Wang, Y., Song, S., Yang, J., Niu, S., Hu, J., Liu, D., Yao, S., Xiong, F., and Li, Z. (2024). Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, 37.

Park, C., Jiang, J., Wang, F., Paul, S., Tang, J., and Kim, S. (2024). Llamaduo: Llmops pipeline for seamless migration from service llms to small-scale local llms. *arXiv preprint arXiv:2408.13467*.

Parnin, C., Soares, G., Pandita, R., Gulwani, S., Rich, J., and Henley, A. Z. (2023). Building your own product copilot: Challenges, opportunities, and needs. *arXiv preprint arXiv:2312.14231*.

Reddy, G. P., Kumar, Y. P., and Prakash, K. P. (2024). Hallucinations in large language models (llms). In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pages 1–6. IEEE.

Respício, A. and Domingos, D. (2015). Reliability of bpmn business processes. *Procedia Computer Science*, 64:643–650.

Roychowdhury, S. (2024). Journey of hallucination-minimized generative ai solutions for financial decision makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1180–1181.

Shan, R. and Shan, T. (2024). Enterprise llmops: Advancing large language models operations practice. In *2024 IEEE Cloud Summit*, pages 143–148. IEEE Computer Society.

Shi, C., Liang, P., Wu, Y., Zhan, T., and Jin, Z. (2024). Maximizing user experience with llmops-driven personalized recommendation systems. *arXiv preprint arXiv:2404.00903*.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D., Horsley, T., Weeks, L., et al. (2018). Prisma extension for scoping reviews (prisma-scr): checklist and explanation. *Annals of internal medicine*, 169(7):467–473.

Venkatapathy, C. (2023). Genai for retail. Technical report, Dell Technologies.

Wang, L. and Zhao, J. (2024). *Strategic Blueprint for Enterprise Analytics*. Springer.