GERENCIAMENTO DO CICLO DE VIDA EM LLMOPS: UM SURVEY SOBRE PRÁTICAS E INTERAÇÕES CONTÍNUAS

LIFECYCLE MANAGEMENT IN LLMOPS: A SURVEY ON PRACTICES AND CONTINUOUS INTERACTIONS

Dirlia Vieira Sousa [1]
José Wellington Franco da Silva [2]
Amanda Drielly Pires Venceslau [3]

RESUMO

Operações com Grandes Modelos de Linguagem (LLMOps) surgem como uma abordagem voltada ao gerenciamento do ciclo de vida de aplicações baseadas em Grandes Modelos de Linguagem (LLMs), abrangendo desde a preparação de dados até a implantação, o monitoramento e a evolução contínua dos modelos em ambientes de produção. Apesar do crescimento recente da área, a literatura ainda se apresenta de forma fragmentada, com estudos que frequentemente abordam essas etapas de maneira isolada ou linear, o que dificulta a compreensão de LLMOps como um processo operacional integrado. Diante desse cenário, este trabalho tem como objetivo analisar, por meio de um *survey* da literatura, como o ciclo de vida de LLMOps é abordado nos estudos existentes, com foco nas práticas e nas interações contínuas entre suas etapas, bem como em sua aplicação em diferentes contextos. A metodologia adotada fundamenta-se em etapas sistemáticas de busca, seleção e análise de estudos primários, seguindo critérios previamente definidos, de modo a assegurar rigor metodológico e consistência na síntese dos resultados. Como resultado, foram selecionados 13 estudos, a partir dos quais foi possível identificar e analisar as principais práticas, processos e interações contínuas que compõem o ciclo de vida de LLMOps, além de compreender sua aplicação em diferentes domínios.

Palavras-chave: LLMOps. LLM. Interação Contínua. Ciclo de Vida de Modelos. Inteligência Artificial.

ABSTRACT

Large Language Model Operations (LLMOps) emerges as an approach focused on managing the lifecycle of applications based on Large Language Models (LLMs), encompassing stages ranging from data preparation to deployment, monitoring, and the continuous evolution of models in production environments. Despite the recent growth of the field, the literature remains fragmented, with studies often addressing these stages in an isolated or linear manner, which hinders a comprehensive understanding of LLMOps as an integrated operational process. In this context, this work aims to analyze, through a literature survey, how the LLMOps lifecycle is addressed in existing studies, with a particular focus on practices and continuous interactions between its stages, as well as on its application across different contexts. The adopted methodology is based on systematic steps for searching, selecting, and analyzing primary studies, following predefined criteria to ensure methodological rigor and consistency in the synthesis of results. As a result, 13 studies were selected, from which it was possible to identify and analyze the main practices, processes, and continuous interactions that comprise the LLMOps lifecycle, as well as to understand its application across different domains.

Keywords: LLMOps. LLM. Continuous Interaction. Model Lifecycle. Artificial Intelligence.

[1] Discente do Curso de Bacharelado em Ciência da Computação da Universidade Federal Do Ceará — Campus Crateús.
[2] Prof. Dr. do Curso de Bacharelado em Ciência da Computação da Universidade Federal do Ceará.
[3] Profa. Dra. da Universidade Federal do Ceará - Instituto Universidade Virtual

1 INTRODUÇÃO

Nos últimos anos, os LLMs passaram a ser amplamente incorporados em aplicações reais, o que intensificou a necessidade de sua operacionalização em ambientes de produção. Todavia, a adoção desses modelos introduz desafios técnicos e operacionais significativos, associados principalmente ao grande volume de dados necessário para treinamento e ajuste, aos elevados custos computacionais e à complexidade de sua manutenção ao longo do tempo (DIAZ-DE-ARCAYA *et al.*, 2024). Em vista disso, práticas manuais tornam-se inviáveis, exigindo abordagens sistemáticas para o gerenciamento do ciclo de vida de aplicações baseadas em LLMs.

Diante desse cenário, as práticas tradicionais de Operações com Modelos de Aprendizagem de Máquina (MLOps), embora forneçam uma base importante para a operacionalização de modelos de aprendizado de máquina, mostram-se insuficientes para lidar com as particularidades dos modelos generativos em larga escala, como apontado por Sinha *et al.* (2024). Como resposta a essas limitações, surge o conceito de LLMOps, abrangendo desde a preparação de dados e o ajuste dos modelos até sua implantação, monitoramento e evolução contínua.

Em contrapartida, a ausência de práticas estruturadas de LLMOps pode resultar em dificuldades relacionadas ao gerenciamento de custos e à utilização eficiente de recursos computacionais. Conforme descrito por Krishnamurthy e Neelanath (2025), a implementação de LLMOps é essencial para gerenciar de forma eficiente custos e recursos no uso de LLMs, especialmente em aplicações de automação inteligente, que demandam alto consumo de recursos computacionais. Deste modo, os autores ressaltam que a gestão eficiente desses recursos contribui para que as aplicações permaneçam econômicas e escaláveis.

Nesse contexto, diferentes estudos no âmbito de LLMOps destacam que o seu ciclo de vida não deve ser compreendido como uma sequência linear de etapas isoladas, mas como um processo marcado por interações contínuas entre suas fases. Shan e Shan (2024) organizam o ciclo de vida de LLMOps em quatro grandes estágios e enfatizam que, ao longo de todo esse processo, a colaboração, a avaliação e a governança contínuas desempenham um papel central para garantir que os modelos de linguagem permaneçam eficazes, éticos e alinhados aos objetivos e valores organizacionais. Esse panorama reforça a ideia de que tais elementos transversais não se restringem a momentos específicos do ciclo, mas permeiam continuamente sua execução.

Em alinhamento com essa visão, Polyakovska (2025) propõe um modelo operacional no qual o gerenciamento do ciclo de vida de LLMs é estruturado como um ciclo fechado, composto por etapas inter-relacionadas. Essa organização cíclica enfatiza a natureza iterativa dos processos envolvidos, favorecendo a melhoria contínua dos modelos em ambientes de produção. De forma complementar, Shi *et al.* (2024) descrevem o LLMOps a partir de fluxos integrados que interagem de maneira contínua, nos quais dados de uso e *feedback* dos usuários alimentam o monitoramento, geram novos insumos para treinamento e ajuste fino, e retornam à implantação e avaliação do modelo. Em conjunto, essas abordagens reforçam a compreensão de que as etapas do ciclo de vida do LLMOps não atuam de forma isolada, mas se influenciam mutuamente por meio de interações contínuas, evidenciando a importância de analisar essas relações ao longo de todo o processo operacional.

Para além da estrutura cíclica, a literatura também indica que as práticas de LLMOps podem assumir diferentes graus de complexidade e relevância ao longo do ciclo de vida dos modelos, dependendo do contexto de aplicação. Mahr *et al.* (2024) destacam que a integração de LLMs em ambientes industriais envolve desafios significativos, especialmente em fases específicas do processo, como a implantação, que tende a apresentar maior complexidade quando comparada a outras etapas. De forma complementar, Chau e Xu (2025) discute a aplicação de LLMs em diversos domínios de atuação, evidenciando que, embora o LLMOps seja essencial

para garantir a eficácia desses modelos em contextos reais, nem todas as etapas do ciclo de vida se manifestam da mesma maneira em cada cenário. Esses estudos evidenciam que as etapas do LLMOps podem apresentar diferentes níveis de relevância e complexidade conforme o domínio de aplicação.

Diante desse cenário, este trabalho propõe um *survey* com o objetivo de analisar a literatura existente sobre LLMOps, com foco na compreensão de seu ciclo de vida como um processo operacional marcado por interações contínuas entre suas etapas. A partir de uma revisão e organização das contribuições atuais, busca-se identificar como essas interações são caracterizadas pelos diferentes autores e de que forma as etapas do ciclo de vida assumem distintos graus de relevância conforme o contexto de aplicação. Ao adotar essa perspectiva integrada, o estudo pretende contribuir para uma visão mais sistemática do LLMOps, oferecendo subsídios conceituais que auxiliem na compreensão de sua dinâmica operacional em ambientes de produção.

1.1 Justificativa

Apesar do recente crescimento das pesquisas em LLMOps, a literatura ainda carece de uma visão consolidada, com estudos majoritariamente voltados à proposição de definições, *frameworks* ou boas práticas de forma isolada. Essa abordagem dificulta a compreensão de LLMOps como um processo operacional integrado e coeso. Muitos trabalhos descrevem o ciclo de vida do LLMOps como uma sequência de etapas, sem aprofundar a análise das interações contínuas e dos mecanismos de retroalimentação que caracterizam sua natureza cíclica, limitando a compreensão de como decisões tomadas em uma etapa influenciam diretamente as demais. Soma-se a isso o fato de que as práticas de LLMOps são aplicadas em contextos diversos, cujas exigências variam conforme o domínio de aplicação, havendo ainda escassez de estudos que analisam de forma sistemática como as etapas do ciclo de vida assumem diferentes graus de relevância nesses contextos. Diante disso, este trabalho justifica-se pela necessidade de organizar e sintetizar a literatura existente, considerando tanto as interações contínuas entre as etapas quanto sua aplicação em diferentes domínios.

1.2 Objetivos

O presente estudo tem como objetivo principal analisar e sistematizar a literatura sobre LLMOps, com foco no ciclo de vida dos LLMs, investigando suas interações contínuas e a aplicação de suas etapas em diferentes domínios.

Os objetivos específicos são os seguintes:

- Levantar e organizar, através de um *survey*, as principais abordagens sobre o ciclo de vida de LLMOps na literatura;
- Avaliar como as etapas do ciclo de vida de LLMOps se conectam e interagem entre si;
- Comparar a relevância e complexidade das etapas do ciclo de vida de LLMOps em diferentes domínios de aplicação;
- Sintetizar os resultados da literatura, destacando pontos de convergência, divergência e lacunas de pesquisa.

1.3 Organização do Trabalho

Este trabalho está dividido em 6 seções. A seção 2 apresenta os fundamentos teóricos relacionados aos LLMs e ao conceito de LLMOps, contextualizando suas principais definições e

características. A seção 3 apresenta os trabalhos relacionados, analisando abordagens existentes e destacando suas contribuições e limitações em relação ao escopo desta pesquisa. A seção 4 descreve a metodologia adotada para a condução deste estudo, incluindo os critérios de seleção e análise dos trabalhos considerados. A seção 5 apresenta os resultados obtidos a partir da análise realizada. Por fim, a seção 6 apresenta as conclusões do trabalho, bem como sugestões para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta a fundamentação teórica que sustenta o desenvolvimento deste trabalho, abordando os principais conceitos relacionados aos LLMs e a LLMOps. São discutidas suas definições, características, formas de utilização e desafios, com o objetivo de estabelecer a base conceitual necessária para a análise e interpretação dos resultados apresentados posteriormente.

2.1 Grandes Modelos de Linguagem

Os Grandes Modelos de Linguagem (LLMs) são uma mudança de paradigma na inteligência artificial, sendo definidos como *foundation models*, ou seja, modelos treinados em larga escala que aprendem um núcleo comum de conhecimentos, podendo ser adaptados a diferentes tarefas com pouca ou nenhuma necessidade de treinamento específico adicional (BOMMASANI *et al.*, 2021). Segundo a definição apresentada pela IBM (2023), esses modelos são baseados em arquiteturas de redes neurais profundas e treinados com extensos conjuntos de dados textuais, o que lhes permite realizar tarefas como compreensão e geração de texto, por meio da modelagem estatística da linguagem.

Essa capacidade de generalização está diretamente relacionada à escala e à arquitetura empregadas na construção desses modelos. Um dos aspectos centrais que definem os LLMs é sua escala. O GPT-3, por exemplo, possui 175 bilhões de parâmetros (BROWN *et al.*, 2020). Esses modelos são treinados em grandes conjuntos de dados, frequentemente contendo bilhões de palavras provenientes de diversas fontes, o que possibilita a geração de conteúdo e previsões por meio da identificação de padrões estatísticos (LIU *et al.*, 2024). Esse nível de complexidade tornou-se viável principalmente com a arquitetura *Transformer*, proposta por Vaswani *et al.* (2017), que se consolidou como base da maioria dos LLMs modernos devido à sua eficiência no tratamento de dependências contextuais em sequências textuais.

Além das características estruturais e arquiteturais, outro elemento relevante diz respeito à forma como esses modelos são disponibilizados e utilizados. Segundo Liu *et al.* (2024), os LLMs podem ser categorizados quanto ao seu modelo de disponibilização, sendo divididos em modelos proprietários (ou de código fechado), como *ChatGPT*, *Claude* e *Gemini*, e modelos de código aberto, como *LLaMA*, *Mistral* e *Falcon*. Estes últimos tendem a oferecer maior transparência e flexibilidade para customização e experimentação, especialmente em contextos de pesquisa.

Apesar de seu desempenho expressivo em diversas tarefas, os LLMs apresentam desafios relevantes. Entre eles destacam-se a geração de informações imprecisas ou inexistentes, conhecida como "alucinação", e a reprodução de vieses e conteúdos tóxicos presentes nos dados de treinamento. Por aprenderem a partir de grandes volumes de textos de múltiplas fontes, esses modelos podem incorporar e amplificar distorções (PATIL; GUDIVADA, 2024). Esses desafios evidenciam que, para além do avanço técnico, o uso de LLMs demanda, portanto, estratégias estruturadas de monitoramento, governança e avaliação contínua.

2.2 Operações com Grandes Modelos de Linguagem

As Operações com Grandes Modelos de Linguagem referem-se a um conjunto de práticas e metodologias voltadas à operacionalização de modelos de linguagem de grande porte em aplicações reais. Esse conceito surge com o objetivo de enfrentar desafios relacionados à implementação desses modelos, como a necessidade de melhorar a precisão, reduzir a latência e otimizar a experiência do usuário (PAHUNE; AKHTAR, 2025). Assim, o LLMOps busca garantir o funcionamento eficiente desses modelos em ambientes de produção, promovendo maior confiabilidade e desempenho.

No que se refere ao escopo de atuação do LLMOps, Shan e Shan (2024) definem essa abordagem como um conjunto de práticas voltadas à gestão operacional do ciclo de vida de modelos de linguagem de grande porte. Nesse sentido, o LLMOps abrange desde etapas iniciais, como o desenvolvimento e treinamento dos modelos, até fases posteriores, como implantação, monitoramento e manutenção em ambientes de produção.

Diante desse cenário, o LLMOps pode ser compreendido como uma adaptação do MLOps, desenvolvida para lidar com os desafios específicos associados ao uso de LLMs (DIAZ-DE-ARCAYA *et al.*, 2024). Esses modelos apresentam características particulares, como o grande volume de dados e a complexidade de sua operacionalização, o que dificulta sua gestão manual e torna sua aplicação distinta das soluções tradicionais de Inteligência Artificial (IA). Embora o MLOps forneça princípios fundamentais para o gerenciamento de modelos, sua aplicação direta aos LLMs mostra-se limitada. Nesse sentido, o LLMOps surge como uma abordagem voltada especificamente às demandas dos modelos generativos, considerando suas particularidades operacionais e viabilizando de forma mais eficiente sua utilização em ambientes corporativos (SINHA *et al.*, 2024).

Apesar dos avanços viabilizados pelo LLMOps, sua aplicação prática ainda enfrenta diversos desafios. Conforme Krishnamurthy e Neelanath (2025), o gerenciamento de dados, uma etapa fundamental do ciclo de vida dos LLMs, envolve atividades como coleta, processamento e armazenamento de grandes volumes de dados, além da necessidade de garantir sua segurança. A etapa de monitoramento de modelos, por sua vez, exige uma infraestrutura robusta, em função do tamanho e da complexidade desses modelos. Adicionalmente, a escala dos LLMs, que pode variar de centenas de milhões a trilhões de parâmetros, intensifica a demanda por recursos computacionais avançados.

Portanto, o LLMOps configura-se como uma evolução das práticas tradicionais de MLOps, ao adaptá-las às exigências específicas dos modelos de linguagem de larga escala. Sua adoção possibilita maior controle, segurança e eficiência ao longo das etapas do ciclo de vida dos LLMs. Contudo, sua aplicação ainda envolve desafios que demandam estratégias especializadas para sua operacionalização, aspectos que se refletem diretamente nas diferentes formas de adoção do LLMOps nos domínios analisados.

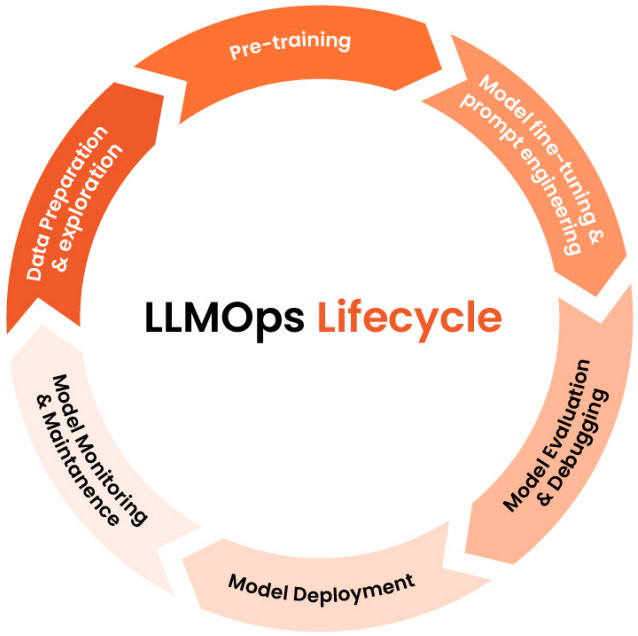
2.3 Ciclo de Vida em LLMOps

O ciclo de vida de LLMOps pode ser definido como o conjunto estruturado de estágios interconectados que orientam o desenvolvimento, a implantação e a manutenção contínua de aplicações baseadas em LLMs em ambientes de produção (Microsoft, 2025). Diferentemente de sistemas de *software* tradicionais, o LLMOps adota uma abordagem iterativa baseada em ciclos de realimentação, reconhecendo que modelos de linguagem exigem aperfeiçoamento contínuo a partir de dados e interações reais (XU *et al.*, 2025).

De modo geral, a literatura descreve o ciclo de vida de LLMOps como uma sequência

estruturada de etapas que operacionalizam as práticas discutidas anteriormente. Essas etapas incluem a preparação dos dados, o treinamento, a avaliação, a implantação e o monitoramento dos modelos (Microsoft, 2025). Em conjunto, essas fases representam o fluxo operacional necessário para levar um modelo de linguagem do desenvolvimento ao ambiente de produção e mantê-lo em funcionamento ao longo do tempo. A Figura 1 apresenta uma representação desse ciclo de vida. Destaca-se que diferentes estudos podem apresentar variações na definição e organização dessas etapas.

Figura 1 – Representação do ciclo de vida do LLMOps.



Fonte: Tredence (2025)

Além dos aspectos técnicos, a literatura recente também destaca a importância de incorporar dimensões de governança e segurança ao ciclo de vida do LLMOps, incluindo práticas voltadas à mitigação de riscos, como ataques de *prompt injection* e o monitoramento de vieses nos modelos (RAVINDRAN, 2025; RAZA *et al.*, 2025).

2.4 Interações Contínuas

Na literatura, a interação contínua entre etapas do ciclo de vida de modelos é frequentemente descrita por meio de conceitos como *feedback loops* e ciclos iterativos. Esses mecanismos permitem a circulação de informações entre diferentes fases do processo. À vista disso, os ciclos de *feedback* possibilitam que resultados obtidos em etapas posteriores, como avaliação ou monitoramento do sistema em operação, sejam incorporados novamente às fases de desenvolvimento (KREUZBERGER *et al.*, 2023). Dessa forma, tornam possível a adaptação do sistema e a realização de ajustes ao longo do ciclo de vida.

Nesse contexto, modelos computacionais passam por diferentes etapas interdependentes, como desenvolvimento, implantação, monitoramento e manutenção. Essas fases não

ocorrem de forma estritamente linear, mas são conectadas por mecanismos que permitem a interação entre diferentes fases do processo. Os *feedback loops*, por sua vez, estabelecem conexões entre as etapas, permitindo que resultados obtidos em uma fase influenciem atividades em outras e possibilitando a avaliação contínua do desempenho dos sistemas em operação. Por meio desses mecanismos, torna-se possível analisar o comportamento do sistema durante sua execução e realizar ajustes quando necessário (THOMPSON, 2024). Como resultado, o ciclo de vida assume um caráter iterativo, no qual as etapas se relacionam continuamente para promover melhorias e adaptações ao longo do tempo diante de mudanças nos dados ou no ambiente de operação.

Diante do exposto, o processo de desenvolvimento e operação de modelos caracteriza-se por interações contínuas entre diferentes fases. Essas interações geralmente são viabilizadas por mecanismos de *feedback*, nos quais informações provenientes de etapas posteriores - como avaliação ou monitoramento - retornam para fases anteriores, como desenvolvimento ou ajuste do modelo, permitindo sua melhoria e adaptação ao longo do tempo.

3 TRABALHOS RELACIONADOS

Nesta seção, são apresentados os trabalhos relacionados que embasam esta pesquisa. A organização desta seção é a seguinte: a Seção 3.1 discute um *survey* técnico sistemático sobre ciclo de vida, ferramentas, desafios e práticas emergentes em LLMOps. A Seção 3.2 explora definições, desafios e o gerenciamento do ciclo de vida desses modelos. A Seção 3.3 analisa propostas de padronização de processos de LLMOps em ambientes corporativos, enquanto a Seção 3.4 apresenta um *framework* prático para a operacionalização de LLMs de forma eficiente.

3.1 Survey técnico sistemático sobre LLMOps: ciclo de vida, ferramentas, desafios e práticas emergentes

Özer (2025) apresenta um *survey* técnico sistemático sobre LLMOps, com o objetivo de analisar o ciclo de vida desses modelos, identificar seus principais desafios operacionais e avaliar estratégias de mitigação associadas à sua aplicação.

A metodologia adotada baseia-se na análise de literatura acadêmica e publicações da indústria, visando compreender as práticas e desafios do LLMOps. A busca foi realizada em bases como *IEEE Xplore*, *ACM Digital Library*, *Scopus* e *arXiv*, utilizando diferentes termos relacionados ao domínio. O processo de seleção seguiu etapas inspiradas nas diretrizes PRISMA, incluindo identificação, triagem e análise completa dos estudos, resultando em um conjunto final de 117 fontes analisadas.

Os resultados apresentam um modelo estruturado de ciclo de vida do LLMOps composto por seis fases, além de identificar quatro categorias principais de desafios: avaliação, aspectos econômicos, gestão de dados e integração. O estudo também destaca a necessidade de abordagens integradas para o gerenciamento desses modelos.

3.2 Operações com Grandes Modelos de Linguagem (LLMOps): definição, desafios e gerenciamento do ciclo de vida

Diaz-De-Arcaya *et al.* (2024) realiza uma revisão da literatura sobre LLMOps, com o objetivo de analisar sua adoção em ambientes de produção, propor uma definição unificada para o termo e identificar os principais desafios e etapas associados ao uso desses modelos. Para isso, o estudo analisou 88 trabalhos inicialmente identificados, passando por etapas de triagem e elegibilidade, até compor um conjunto final de 34 trabalhos relevantes.

A análise identificou uma definição unificada de LLMOps e os principais desafios relacionados à sua aplicação em ambientes de produção. O estudo também detalha as etapas do ciclo de vida dessa metodologia, evidenciando a necessidade de adaptação de práticas tradicionais como DevOps e MLOps.

3.3 Rumo a um modelo padronizado de processos de negócio para LLMOps

Chernigovskaya *et al.* (2025) propõe um modelo padronizado de processos para LLMOps, com o objetivo de apoiar a adoção e o gerenciamento de modelos de linguagem de larga escala em ambientes corporativos.

A pesquisa adota uma revisão sistemática da literatura, baseada em uma estratégia de busca por palavras-chave e em etapas de seleção que consideram critérios de inclusão e exclusão. Esse processo resultou em um conjunto refinado de estudos analisados.

Assim, a análise revelou que não existem modelos padronizados para LLMOps na literatura. Com base nessa lacuna, o estudo propõe um modelo de processos de negócio, representado em BPMN (*Business Process Model and Notation*), com o objetivo de organizar e padronizar a integração de LLMs em ambientes corporativos.

3.4 Em direção a uma estrutura prática para LLMOps: Compreendendo e construindo a excelência operacional para grandes modelos de linguagem

Polyakovska (2025) explora o conceito de LLMOps sob uma perspectiva prática, com o objetivo de analisar suas particularidades em relação ao MLOps e propor um *framework* para a operacionalização de modelos de linguagem de larga escala.

Neste estudo, adotou-se uma abordagem qualitativa e conceitual, fundamentada na análise de literatura e de práticas contemporâneas envolvendo LLMs, MLOps e LLMOps. O foco está na identificação de desafios técnicos e éticos e na organização das etapas do ciclo de vida desses modelos. Como resultado, é proposto um *framework* prático de LLMOps, estruturado em fases do ciclo de vida, que inclui recomendações de monitoramento, governança e mitigação de riscos, ressaltando a importância de práticas adaptadas às especificidades dos modelos de linguagem de larga escala.

3.5 Análise comparativa

Com o intuito de sistematizar as contribuições dos principais trabalhos analisados e evidenciar suas convergências e limitações, apresenta-se a seguir uma análise comparativa entre os estudos selecionados e a abordagem adotada neste trabalho. A comparação considera critérios relacionados à definição de LLMOps, ao nível de detalhamento das etapas, à consideração de diferentes domínios de aplicação e à forma como as interações entre as etapas são tratadas. Essa análise permite situar a proposta deste estudo em relação ao panorama da literatura, evidenciando suas características distintivas e sua inserção no campo de pesquisa em LLMOps, conforme apresentado na Tabela 1.

A análise comparativa apresentada evidencia que, embora os estudos analisados ofereçam contribuições relevantes para a consolidação do conceito de LLMOps - seja por meio da proposição de *frameworks*, modelos de processos ou revisões sistemáticas da literatura -, ainda existem limitações no que se refere à análise das interações contínuas entre as etapas do ciclo de vida e à consideração aprofundada de múltiplos domínios de aplicação. De modo geral, esses trabalhos concentram-se na definição de estruturas, identificação de desafios e proposição

Tabela 1 – Síntese dos Trabalhos Relacionados

Critério	**Özer (2025)**	**Diaz-De-Arcaya et al. (2024)**	**Chernigovskaya et al. (2025)**	**Polyakovska (2025)**	**Este trabalho**
Objetivo	Survey sobre ciclo de vida, desafios e práticas de LLMOps	Definição unificada e análise da adoção de LLMOps	Proposição de modelo padronizado de processos (BPMN)	Framework prático para operacionalização de LLMOps	Análise das interações e aplicação das etapas do LLMOps
Análise de domínios de aplicação	Não realiza análise específica por domínio	Não realiza análise por domínio	Foco em processos organizacionais, sem análise por domínio	Apresenta exemplos práticos, sem foco comparativo	Analisa a aplicação das etapas em diferentes domínios
Interações entre etapas	Não analisa	Não analisa explicitamente	Aborda fluxo de processos, mas sem foco em interações contínuas	Não analisa	Analisadas como processos inter-relacionados no ciclo de vida
Abordagem do ciclo de vida	Modelo estruturado com fases e desafios associados	Identificação e descrição das etapas do ciclo de vida	Modelo de processos de negócio para LLMOps	Modelo prático e aplicado	Visão analítica e comparativa baseada na literatura

Fonte: Elaborado pelo autor (2026)

de boas práticas, sem, contudo, explorar de forma integrada como essas etapas se articulam e se adaptam a diferentes contextos. Diante disso, a abordagem adotada neste estudo busca ampliar essa perspectiva ao investigar tais lacunas, contribuindo para uma compreensão mais abrangente e inter-relacionada do LLMOps.

4 METODOLOGIA

Nesta seção é apresentada a metodologia adotada para a elaboração deste estudo. São descritos os procedimentos utilizados para estruturar a pesquisa, incluindo a definição dos objetivos, o processo de coleta dos artigos, as estratégias de busca empregadas e os critérios de inclusão e exclusão aplicados durante a seleção do material analisado. Dessa forma, esta seção estabelece a base necessária para sustentar a análise apresentada nas etapas seguintes do trabalho.

4.1 Objetivos da pesquisa

As questões que guiam este *survey* estão relacionadas ao gerenciamento do ciclo de vida do LLMOps. Como mencionado por Diaz-De-Arcaya *et al.* (2024), o termo é tão novo que a literatura científica ainda não chegou a um acordo sobre sua própria definição. Dessa forma, ainda há pouca consolidação teórica sobre esse assunto, sobretudo com relação a suas fases e interações contínuas entre elas. Apesar dessas fases serem citadas em muitos artigos, sua ocorrência é de forma bastante dispersa, o que dificulta o entendimento acerca do tema.

Sendo assim, este *survey* tem como objetivo geral organizar, sintetizar e analisar informações, de modo a apoiar estudos futuros, fornecer uma base conceitual e mostrar como LLMOps está sendo aplicado em diferentes áreas. Para documentar todo o procedimento, utilizou-se o *Parsifal*, uma ferramenta *online* criada para auxiliar na realização de revisões sistemáticas. A ferramenta segue as diretrizes de Kitchenham e Charters (2007). O processo dispõe de três etapas principais: planejamento, condução e relatório.

4.2 Planejamento

Na etapa de planejamento, é elaborado um protocolo de revisão que define as questões de pesquisa, a estratégia de busca, as bases de dados e os critérios de seleção. Assim, os principais aspectos do protocolo serão apresentados a seguir.

4.2.1 Questões de Pesquisa

Foram definidas duas questões a serem respondidas:

- QP1: Quais são as interações contínuas entre as etapas do LLMOps?
- QP2: Como as etapas estão sendo aplicadas em diferentes domínios?

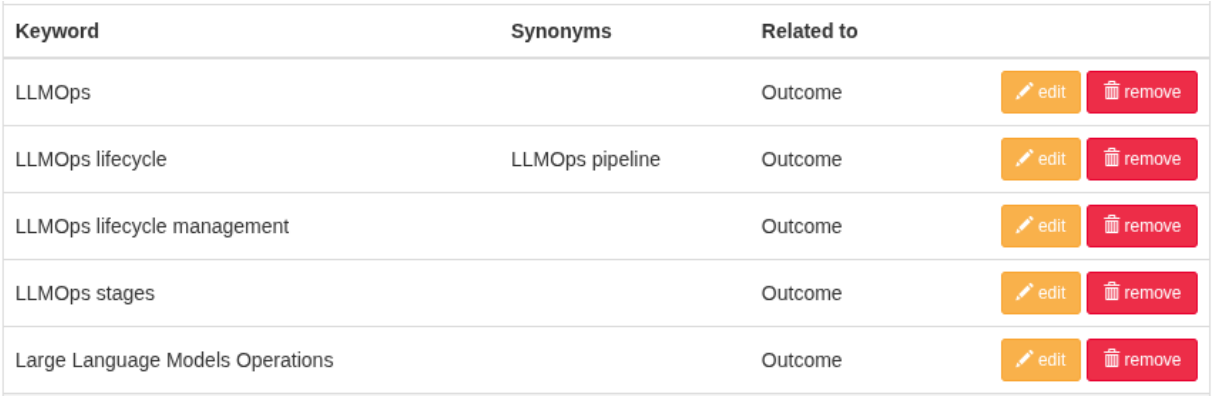
4.2.2 Estratégia de Busca

As estratégias de busca foram definidas com o objetivo de garantir uma recuperação ampla e ao mesmo tempo precisa dos estudos relacionados ao tema de LLMOps. Para isso, foram combinados palavras-chave, operadores booleanos e sinônimos diretamente associados aos conceitos centrais da pesquisa.

Por meio do protocolo PICO (*Population*, *Intervention*, *Comparison*, *Outcomes*) (RICHARDSON *et al.*, 1995), que advém dos elementos relacionados a paciente, problema ou população (P), intervenção (I), comparação, controle ou comparador (C), e resultado (O), e das palavras-chave, formadas a partir da combinação de sinônimos desses elementos, foi elaborada a *string* de busca.

Primeiro, foram identificados os principais termos presentes na literatura, incluindo expressões como "LLMOps", "Large Language Model Operations", "LLMOps Lifecycle", "LLMOps Lifecycle management", "LLMOps stages". Esses termos também foram combinados com sinônimos e variações comuns na literatura, de modo a aumentar a abrangência dos resultados. A definição e a organização dessas palavras-chave são ilustradas na Figura 2.

Figura 2 – Definição e organização das palavras-chave utilizadas na pesquisa na plataforma Parsifal.



Fonte: Elaborado pela autora (2026)

Em seguida, foram construídas expressões booleanas utilizando os operadores AND e OR. O objetivo foi conectar conceitos complementares e evitar a recuperação de estudos irrelevantes. A string de busca final foi adaptada conforme as particularidades de cada base de dados consultada, uma vez que IEEE Xplore, ACM Digital Library e Scopus possuem diferenças em seus mecanismos de filtragem e suporte a operadores.

A principal string de busca adotada neste estudo é apresentada na Figura 3.

Figura 3 – String de busca utilizada para a seleção dos estudos.

Base String

```
(LLMOps OR "Large Language Model Operations") AND (Lifecycle OR "lifecycle management" OR stages)
```

Fonte: Elaborado pela autora (2026)

4.3 Fontes de Pesquisa

Para a obtenção de dados, foram selecionadas três fontes de busca, as quais estão listadas a seguir:

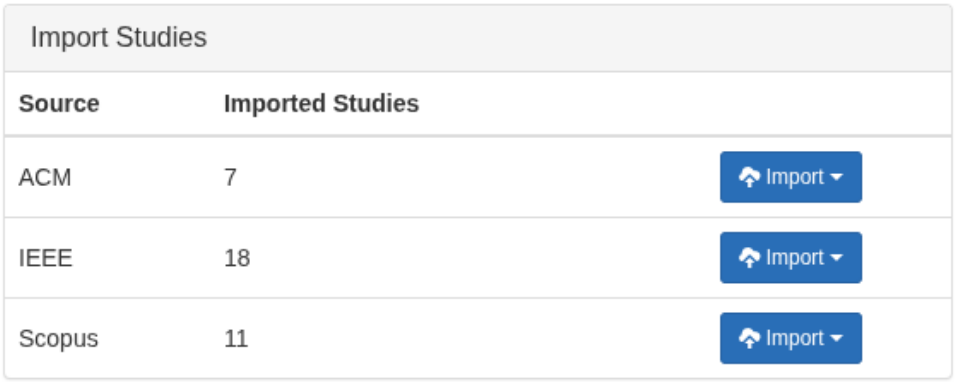
- *ACM Digital Library*[4];
- *IEEE Xplore Digital Library* [5] e;
- *Scopus* [6].

As fontes em questão foram escolhidas por se tratarem das principais bibliotecas digitais na área da computação e pela ampla quantidade de estudos pertinentes ao tema disponíveis nesses repositórios.

Embora não tenha sido estabelecido um recorte temporal prévio, os estudos selecionados concentram-se nos anos de 2024 e 2025, o que reflete o caráter recente do tema e a disponibilidade de publicações na área.

Após a etapa de busca, os estudos recuperados foram importados para a plataforma *Parsifal*, que foi utilizada para organizar os registros, remover duplicatas e estruturar as etapas subsequentes de triagem. A Figura 4 apresenta a interface da plataforma *Parsifal* com os estudos importados após a etapa de busca, destacando a organização dos registros que antecede as fases de triagem e seleção.

Figura 4 – Organização dos estudos importados na plataforma Parsifal após a etapa de busca.



Fonte: Elaborado pela autora (2026)

[4] https://dl.acm.org/
[5] http://ieeexplore.ieee.org/
[6] http://www.scopus.com/

4.4 Critérios de inclusão e exclusão

Após o levantamento preliminar dos artigos a partir das strings de busca citadas na seção anterior, foi realizada uma leitura inicial dos títulos, resumos e, quando necessário, do texto completo, a fim de avaliar a adequação às questões de pesquisa e aplicar os critérios de inclusão e exclusão estabelecidos. Esses critérios foram definidos para garantir que apenas trabalhos relevantes e alinhados ao escopo da pesquisa fossem analisados.

4.4.1 Critérios de Inclusão

- CI1: Artigos completos;
- CI2: Artigos disponíveis nas fontes de pesquisa;
- CI3: Artigos em português ou inglês;
- CI4: Estudos respondem às questões de pesquisa.

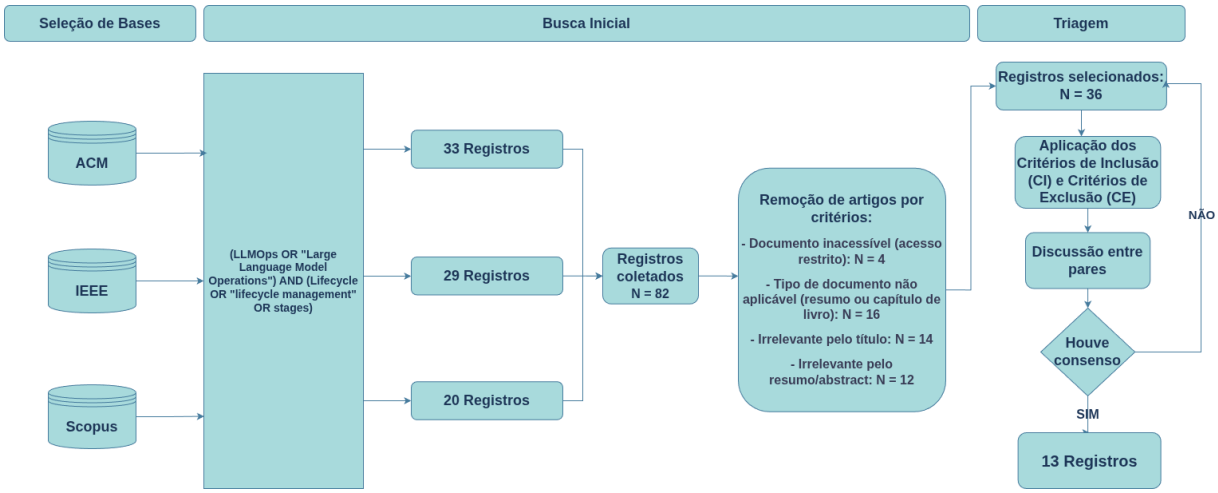
4.4.2 Critérios de Exclusão

- CE1: Artigos com acesso restrito;
- CE2: Artigos duplicados;
- CE3: Estudos não respondem às questões de pesquisa;
- CE4: Idioma que não seja português ou inglês;
- CE5: Não é um estudo primário;
- CE6: Pôsteres, resumo, capítulo de livro ou artigo curto.

4.5 Condução

As etapas para a realização do estudo proposto encontram-se ilustradas na Figura 5.

Figura 5 – Fluxograma do processo metodológico de seleção dos estudos.



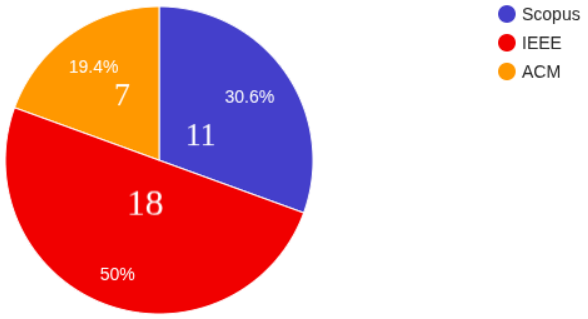
Fonte: Elaborado pela autora (2026)

Na etapa inicial, foi realizada uma triagem preliminar dos estudos com base em critérios de exclusão estabelecidos, resultando na remoção de 46 artigos. Desses, 4 foram excluídos por apresentarem acesso restrito, 16 por não se enquadrarem no tipo de publicação (resumos ou capítulos de livro), 14 por irrelevância identificada a partir da leitura dos títulos e

12 por não atenderem aos critérios após a análise dos resumos (*abstracts*). Após essa etapa, os estudos remanescentes foram submetidos à triagem manual, com leitura completa dos artigos.

A distribuição dos estudos recuperados por base de dados é apresentada na Figura 6, permitindo observar a participação de cada base no conjunto inicial de artigos.

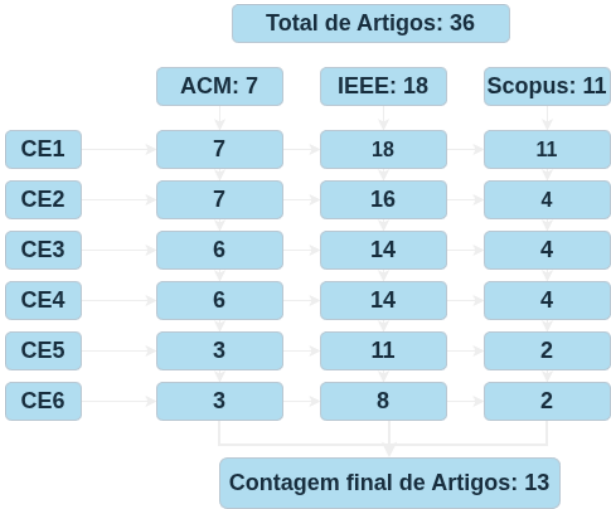
Figura 6 – Distribuição dos artigos analisados



Fonte: Elaborado pela autora (2026)

A Figura 7 apresenta o processo de triagem manual dos estudos, evidenciando a redução progressiva do número de artigos em cada base de dados (ACM, IEEE e Scopus) a partir da aplicação dos critérios de exclusão (CE1–CE6). Inicialmente, foram identificados 36 estudos, que foram filtrados sequencialmente conforme os critérios definidos. Ao final desse processo, obteve-se um total de 13 artigos selecionados para análise. Os critérios de exclusão adotados neste estudo estão descritos previamente na Seção 4.4.2.

Figura 7 – Fluxograma do processo de triagem dos artigos.

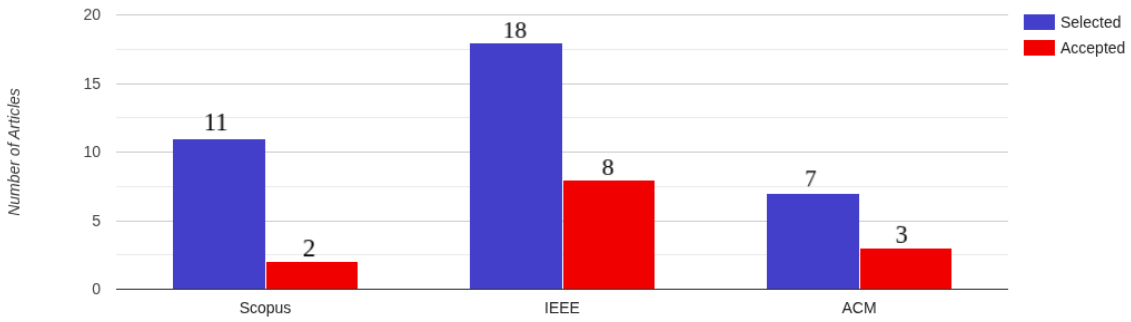


Fonte: Elaborado pela autora (2026)

Ao final das etapas de filtragem, um total de 13 estudos permaneceu elegível para a análise. A distribuição de artigos aceitos foi a seguinte: pelo IEEE Xplore (8), pelo ACM Digital Library (3) e pelo Scopus (2), conforme ilustrado no gráfico da figura 8. Esses artigos compõem o conjunto final a partir do qual os resultados dessa pesquisa foram estruturados. Portanto, a

apresentação desses dados fornece um panorama geral do material selecionado e estabelece a base quantitativa que sustenta as análises apresentadas nas subseções seguintes.

Figura 8 – Distribuição dos artigos aceitos após análise.



Fonte: Elaborado pela autora (2026)

A Tabela 2 apresenta os estudos selecionados após o processo de triagem. Esses estudos compõem o conjunto final analisado neste trabalho, servindo como base para a investigação das práticas, etapas e interações associadas ao ciclo de vida do LLMOps.

Tabela 2 – Síntese dos estudos selecionados

Referência
Abdellatif *et al.* (2025)
Bodor *et al.* (2025)
Chau e Xu (2025)
Chen *et al.* (2025)
Gupta *et al.* (2024)
Krishnamurthy e Neelanath (2025)
Mahr *et al.* (2024)
Miao *et al.* (2024)
Myakala *et al.* (2025)
Pahune e Akhtar (2025)
Radenkovic *et al.* (2025)
Shan e Shan (2024)
Sinha *et al.* (2024)

Fonte: Elaborado pela autora (2026)

5 RESULTADOS

Esta seção apresenta os resultados obtidos a partir da análise dos estudos selecionados. Com o objetivo de responder às questões de pesquisa definidas neste trabalho, os resultados foram organizados de acordo com cada uma das questões propostas.

5.1 QP1: Quais são as interações contínuas entre as etapas do LLMOps?

A primeira questão de pesquisa busca identificar as interações contínuas entre as etapas do LLMOps. Nesse contexto, são descritas as principais etapas identificadas nos estudos, bem como as interações estabelecidas entre elas.

5.1.1 Principais etapas do LLMOps

Os artigos analisados apresentam, de maneira geral, um conjunto de etapas semelhantes, conforme sintetizado na Tabela 3. Dentre elas, destacam-se: Preparação dos Dados, Treinamento, Avaliação, Implantação e Monitoramento. Observa-se que a maioria dos estudos adota uma estrutura semelhante para descrever o ciclo de vida do LLMOps. No entanto, há pequenas variações na nomenclatura dessas etapas, ainda que desempenhem funções equivalentes. A etapa de Preparação dos Dados, por exemplo, recebe denominações como Pré-processamento e Gerenciamento de Dados, conforme apresentado nos trabalhos de Sinha *et al.* (2024) e Krishnamurthy e Neelanath (2025). Apesar dessas variações, as funções desempenhadas permanecem essencialmente as mesmas.

5.1.2 Interações contínuas entre as etapas do LLMOps

Além das etapas que compõem o ciclo de vida do LLMOps, os estudos analisados evidenciam que elas não operam de forma isolada. Pelo contrário, há interações contínuas e cíclicas entre elas, que sustentam o funcionamento iterativo do LLMOps.

A etapa de preparação de dados envolve o tratamento de informações provenientes de diferentes fontes, como repositórios de código e documentação. Durante esse processo, são definidos os casos de uso e realizado o planejamento do pipeline de dados, incluindo a identificação das fontes e o fluxo de informações que alimentarão o sistema (SINHA *et al.*, 2024; RADENKOVIC *et al.*, 2025).

Nesse contexto, os dados passam inicialmente por processos de limpeza, remoção das duplicatas e filtragem, garantindo a qualidade do conjunto de dados e a remoção de conteúdos indesejados, como informações tóxicas ou conflitos de licenciamento. Em seguida, são convertidos para formatos compatíveis com os modelos de linguagem, sendo organizados em blocos de texto e posteriormente divididos em unidades menores, denominadas *tokens*, por meio de um processo de *tokenização*. Por fim, esses *tokens* são transformados em representações numéricas, como vetores, possibilitando sua utilização no treinamento dos modelos (MAHR *et al.*, 2024; RADENKOVIC *et al.*, 2025). Esse conjunto de ações contribui para que os dados estejam estruturados e relevantes, estabelecendo uma base adequada para a etapa de treinamento (SHAN; SHAN, 2024). Nessa perspectiva, a etapa de treinamento do modelo depende diretamente da qualidade dos dados provenientes do pré-processamento, bem como das configurações previamente definidas, sendo ajustado com base nessas informações, o que resulta em melhorias no desempenho e na precisão das respostas geradas (GUPTA *et al.*, 2024).

Como resultado do processo de treinamento, obtém-se um modelo ajustado, cujo desempenho serve como base para a etapa de avaliação (SINHA *et al.*, 2024; SHAN; SHAN, 2024). Nessa etapa, o modelo é analisado por meio de métricas de desempenho, como o erro quadrático médio (MSE) e o coeficiente de determinação (R^2), com o objetivo de verificar sua eficácia e a precisão dos resultados gerados (BODOR *et al.*, 2025). Além disso, a avaliação pode envolver o uso de *benchmarks* e testes específicos, como avaliações de alucinação e a aplicação de *prompts* adversariais, possibilitando uma análise mais abrangente da qualidade

e confiabilidade das respostas geradas. Esses métodos permitem identificar possíveis falhas e limitações do modelo em diferentes cenários (RADENKOVIC *et al.*, 2025).

Com base nessa análise, são realizados processos de ajuste fino, que podem envolver modificações na arquitetura do modelo, refinamento de hiperparâmetros ou a reexecução do treinamento com novos dados. Essa interação entre treinamento, avaliação e ajuste fino contribui para o aprimoramento progressivo do modelo (GUPTA *et al.*, 2024). À vista disso, observa-se que as etapas de treinamento e avaliação estão inseridas em um ciclo contínuo de refinamento, no qual os resultados obtidos orientam ajustes nos dados e nos parâmetros do modelo, promovendo a melhoria contínua de seu desempenho ao longo do tempo (MIAO *et al.*, 2024).

Na etapa de avaliação, por sua vez, os resultados obtidos desempenham um papel fundamental na decisão de implantação do modelo, uma vez que permitem verificar se os critérios de desempenho e segurança foram atendidos. Dessa forma, apenas modelos que apresentam resultados satisfatórios são encaminhados para ambientes de produção, garantindo maior confiabilidade e qualidade nas aplicações desenvolvidas (RADENKOVIC *et al.*, 2025). Ademais, a avaliação do modelo caracteriza-se como um processo iterativo, no qual diferentes métodos são empregados para analisar seu desempenho e eficácia. Após a validação, o modelo pode ser versionado e armazenado em repositórios específicos, possibilitando o acompanhamento de sua evolução ao longo do tempo e contribuindo para uma transição mais estruturada para o ambiente de produção (MAHR *et al.*, 2024).

Uma vez definido que o modelo atende aos critérios estabelecidos, a etapa de implantação é responsável por disponibilizá-lo em ambientes de produção, garantindo seu uso de forma prática e em escala. Diante disso, a escolha da estratégia de implantação deve considerar fatores como orçamento, requisitos de segurança e a infraestrutura disponível, de modo a assegurar uma operação eficiente e confiável, conforme apontado por Sinha *et al.* (2024).

Após a implantação, os modelos são frequentemente containerizados e disponibilizados por meio de APIs, sendo integrados a *pipelines* de CI/CD para viabilizar sua utilização em aplicações reais. Essa estrutura não apenas facilita o acesso ao modelo em produção, mas também estabelece uma base adequada para o monitoramento contínuo de seu desempenho (RADENKOVIC *et al.*, 2025). A partir disso, inicia-se o monitoramento contínuo do modelo em ambiente de produção, com o objetivo de acompanhar seu desempenho ao longo do tempo. Nesse processo, são coletadas métricas, identificados erros e analisado o comportamento do sistema em cenários reais, permitindo identificar possíveis falhas, degradação de desempenho ou comportamentos inesperados (PAHUNE; AKHTAR, 2025; CHEN *et al.*, 2025).

A partir do monitoramento contínuo, são coletadas informações sobre o desempenho do sistema em uso real, permitindo a identificação de possíveis falhas ou oportunidades de melhoria. Ao longo desse processo, são obtidos tanto dados relacionados ao desempenho do modelo em cenários reais quanto *feedback* proveniente dos usuários, possibilitando uma análise mais abrangente do comportamento do sistema (KRISHNAMURTHY; NEELANATH, 2025; CHEN *et al.*, 2025). Durante essa etapa, os dados coletados em cenários reais de uso do modelo são utilizados para retroalimentar o ciclo de vida do LLM, possibilitando a identificação de desvios ou degradação de desempenho (SHAN; SHAN, 2024). Com base nessas informações, são realizados ajustes e processos de retreinamento ou refinamento do modelo, com o objetivo de aprimorar seu desempenho, conectando diretamente o monitoramento às etapas de treinamento (SINHA *et al.*, 2024), configurando um monitoramento contínuo essencial para garantir a precisão e a confiabilidade dos resultados ao longo do tempo. Esse processo estabelece um *feedback loop* que contribui para a adaptação e melhoria contínua do sistema (BODOR *et al.*, 2025; CHEN *et al.*, 2025).

A fim de sintetizar as interações contínuas identificadas entre as etapas do LLMOps,

a Tabela 3 apresenta uma visão estruturada das principais relações observadas, evidenciando como as etapas se conectam de forma contínua ao longo do ciclo de vida.

Tabela 3 – Interações contínuas entre as etapas do ciclo de vida do LLMOps

Categoria	**Interação**	**Referências**
Entrada de dados e encaminhamento	Preparação dos dados → Treinamento do modelo	(Bodor et al., 2025; Mahr et al., 2024; Radenkovic et al., 2025; Shan et al., 2024; Sinha et al., 2024)
	Avaliação → Implantação	(Bodor et al., 2025; Chen et al., 2025; Gupta et al., 2024; Mahr et al., 2024; Radenkovic et al., 2025)
	Implantação → Monitoramento	(Chen et al., 2025; Pahune et al., 2025; Radenkovic et al., 2025; Sinha et al., 2024)
Retroalimentação	Monitoramento → Treinamento (retreinamento)	(Abdellatif et al., 2025; Bodor et al., 2025; Chen et al., 2025; Krishnamurthy et al., 2025; Mahr et al., 2024; Shan et al., 2024; Sinha et al., 2024)
Ajuste iterativo	Treinamento ↔ Avaliação	(Gupta et al., 2024; Miao et al., 2024; Radenkovic et al., 2025; Shan et al., 2024; Sinha et al., 2024)

Fonte: Elaborado pela autora (2026)

As interações classificadas como entrada de dados e encaminhamento representam o fluxo inicial e a progressão entre as etapas do ciclo de vida, refletindo o processo de preparação, desenvolvimento e disponibilização dos modelos. Já as interações de retroalimentação evidenciam o retorno de informações provenientes do ambiente de produção para o aprimoramento dos modelos, enquanto as interações de ajuste iterativo indicam processos contínuos de refinamento entre etapas interdependentes.

5.2 QP2: Como as etapas estão sendo aplicados em diferentes domínios?

A aplicação das etapas do LLMOps pode variar de acordo com o domínio em que os modelos são utilizados, uma vez que diferentes contextos impõem requisitos e desafios específicos. Nesse sentido, determinadas etapas do ciclo de vida tendem a assumir maior relevância conforme o cenário de aplicação. Sob essa perspectiva, torna-se importante analisar como essas etapas são aplicadas em diferentes domínios, a fim de compreender suas particularidades e identificar os aspectos que influenciam seu funcionamento em cada área.

5.2.1 Domínios críticos: Saúde, Finanças, Educação e Jurídico

Em domínios críticos e regulados, como saúde e finanças, a aplicação das etapas do LLMOps exige atenção a diferentes requisitos específicos. No contexto da saúde, destaca-se a necessidade de um elevado nível de precisão nos resultados gerados pelos modelos, especialmente em cenários que envolvem a tomada de decisão clínica, nos quais erros decorrentes de alucinações podem comprometer diretamente a qualidade do atendimento. Diante disso, a etapa de monitoramento se torna particularmente crítica, uma vez que o acompanhamento contínuo do desempenho do modelo é essencial para garantir a confiabilidade das respostas e minimizar

a ocorrência de resultados incorretos. Além disso, nesses domínios, aspectos relacionados à governança ganham maior relevância ao longo do ciclo de vida do LLMOps, sobretudo no que se refere à segurança dos dados, privacidade e conformidade regulatória. Esses requisitos impactam diretamente etapas como a implantação e o monitoramento, nas quais mecanismos de auditabilidade e rastreabilidade tornam-se fundamentais para o acompanhamento das saídas, atualizações e interações do modelo, contribuindo para garantir que as decisões automatizadas sejam transparentes e verificáveis (SINHA *et al.*, 2024; KRISHNAMURTHY; NEELANATH, 2025).

De modo semelhante, Krishnamurthy e Neelanath (2025) destacam que, no domínio jurídico, a aplicação das etapas do LLMOps pode incorporar abordagens de *human-in-the-loop*, especialmente nas fases de preparação de dados, treinamento e refinamento do modelo. Em tais casos, a intervenção humana atua na orientação e validação das decisões do sistema, sendo particularmente relevante em cenários de alto risco, nos quais erros podem gerar consequências significativas. Com isso, a integração entre humanos e modelos contribui para aumentar o controle, a confiabilidade e a segurança das aplicações baseadas em LLMs.

De forma complementar, em uma perspectiva mais ampla, o estudo de Chau e Xu (2025) evidencia que, em diferentes domínios de aplicação, como saúde, finanças, educação e jurídico, a adoção de práticas de LLMOps é fundamental para garantir a eficácia dos modelos. Embora nem sempre sejam explicitadas as etapas mais críticas em cada contexto, destaca-se a importância do processo de alinhamento, inserido na etapa de treinamento. Esse processo tem como objetivo evitar a geração de respostas desalinhadas com valores, objetivos e expectativas humanas, contribuindo para uma atuação mais segura e confiável dos modelos em diferentes cenários.

5.2.2 Aplicações corporativas e industriais

No setor empresarial, a eficácia do LLMOps está diretamente ligada à integração com a infraestrutura de TI existente e à capacidade de atender a diferentes demandas de negócio. Nesse sentido, a etapa de preparação de dados envolve o gerenciamento de grandes volumes de informações e sua adaptação a múltiplas aplicações, por meio de *pipelines* estruturados. Ademais, a implantação requer a integração eficiente com sistemas legados, visando minimizar impactos nas operações. Já a etapa de monitoramento e manutenção assume papel relevante na análise de desempenho, escalabilidade e na realização de atualizações contínuas. Por fim, aspectos de governança de dados, como privacidade e segurança, também se destacam, contribuindo para o cumprimento de regulamentações e a confiança dos usuários (SHAN; SHAN, 2024).

Já no setor imobiliário, a aplicação das etapas do LLMOps é marcada pelo uso do ajuste fino e do monitoramento contínuo, especialmente em cenários que envolvem o processamento de dados em tempo real. O ajuste fino permite adaptar os modelos às necessidades específicas do setor, como análise de contratos, avaliação de imóveis e atendimento ao cliente. Por sua vez, o monitoramento contínuo torna-se essencial para garantir a confiabilidade do modelo diante de dados dinâmicos, como variações de preços e mudanças na disponibilidade de imóveis. Assim, essas etapas contribuem para manter o desempenho e a relevância dos modelos em um ambiente caracterizado por constantes atualizações, conforme evidenciado no estudo de Bodor *et al.* (2025).

Em contraste, no setor industrial, a aplicação das etapas do LLMOps apresenta desafios específicos relacionados à infraestrutura e às restrições de segurança, especialmente na etapa de implantação do modelo. Em muitos casos, os dispositivos utilizados em ambientes de produção não possuem acesso à internet, o que exige a adoção de estratégias diferenciadas

para viabilizar o uso dos LLMs. Nesse cenário, uma das abordagens consiste na execução local do modelo, por meio de sua implantação em dispositivos de borda virtual (*vEdge*) ou em computadores industriais com maior capacidade computacional. Alternativamente, pode-se utilizar o *prompting* por meio de chamadas de Interface de Programação de Aplicações (API), permitindo o processamento do modelo de forma remota. (MAHR *et al.*, 2024) Essas estratégias evidenciam a necessidade de adaptação da etapa de implantação às limitações do ambiente industrial.

5.2.3 Sistemas inteligentes e automação

No âmbito da automação inteligente, o ajuste fino do modelo assume papel central, permitindo que os LLMs evoluam continuamente e se adaptem a mudanças nos requisitos de negócio, em ambientes regulatórios ou nas preferências dos usuários. Simultaneamente, o monitoramento contínuo se mostra essencial, garantindo que os processos permaneçam confiáveis e eficientes ao longo do tempo. Essas etapas do LLMOps são fundamentais para assegurar a robustez e a precisão das soluções automatizadas, mantendo a eficácia dos modelos em cenários dinâmicos e complexos (KRISHNAMURTHY; NEELANATH, 2025).

Em ambientes de Internet das Coisas (IoT) e infraestruturas inteligentes, as etapas de ajuste fino e implantação do LLMOps são essenciais para viabilizar a utilização de grandes modelos de linguagem em dispositivos de borda, sensores e sistemas IoT. O ajuste fino permite adaptar os modelos para tarefas específicas de IoT e reduzir custos computacionais, garantindo que os LLMs se ajustem às particularidades do domínio. Já a implantação precisa lidar com limitações de memória, capacidade computacional reduzida e a necessidade de respostas rápidas, desafios típicos desses dispositivos. Desse modo, o ajuste fino e a implantação asseguram que os modelos operem de forma eficiente e confiável nos dispositivos de borda, mantendo desempenho adequado mesmo diante de recursos limitados (MYAKALA *et al.*, 2025).

5.2.4 Processamento de documentos (OCR e reconhecimento de caracteres)

A aplicação de LLMs no reconhecimento de caracteres manuscritos tem otimizado tarefas de processamento de imagens e textos manuscritos, oferecendo maior precisão e velocidade em comparação com métodos tradicionais de Reconhecimento Óptico de Caracteres (OCR). Nesse contexto, as etapas de preparação de dados e de treinamento e ajuste fino do modelo se mostram fundamentais. A preparação de dados envolve a utilização de conjuntos de imagens de texto manuscrito como dados de entrada, garantindo que o modelo receba informações consistentes e de qualidade. Já o treinamento e ajuste fino permitem que os LLMs sejam adaptados para reconhecer padrões específicos de escrita, aumentando a acurácia do reconhecimento (GUPTA *et al.*, 2024).

No processamento automatizado de documentos, a implantação de grandes modelos de linguagem é um desafio crítico, devido ao elevado custo computacional e às limitações de recursos em cenários de processamento em tempo real. É preciso planejar cuidadosamente a infraestrutura e os requisitos do sistema para que os modelos possam operar de forma eficiente e confiável. Em paralelo, a etapa de avaliação assume papel central, permitindo monitorar constantemente a precisão das saídas, identificar erros como alucinações ou interpretações incorretas de campos provenientes de OCR e realizar ajustes finos com base em *feedback*. Como resultado, essas etapas garantem que os modelos mantenham desempenho confiável e que os documentos processados sejam estruturados corretamente e utilizáveis em aplicações reais (ABDELLATIF *et al.*, 2025).

5.2.5 Engenharia de software

No domínio de Engenharia de Software, Radenkovic *et al.* (2025) destacam que os LLMs atuam como assistentes de desenvolvimento, sendo usados para gerar código, criar testes, explicar erros, sugerir correções, auxiliar em *pipelines de CI/CD* e analisar *logs*. Apesar de seu potencial, esses modelos apresentam limitações, como alucinações, respostas incorretas, falhas de lógica, sensibilidade a *prompts* e questões legais e éticas, que podem gerar *bugs*, vulnerabilidades ou falhas em sistemas em produção. É nesse contexto que a etapa de avaliação e validação do LLMOps se mostra essencial, garantindo que as saídas sejam verificadas antes de qualquer uso real. Essa etapa permite identificar erros de lógica, respostas imprevisíveis e potenciais riscos de segurança, promovendo a intervenção humana quando necessário e assegurando a confiabilidade do modelo no processo de desenvolvimento.

5.2.6 Síntese dos domínios e etapas do LLMOps

Com o objetivo de resumir os principais domínios de aplicação do LLMOps e as etapas que mais se destacam em cada contexto, a Tabela 4 apresenta uma visão geral dessas relações, destacando como as diferentes áreas influenciam o comportamento do ciclo de vida dos modelos.

Tabela 4 – Domínios de aplicação e etapas críticas do LLMOps

Domínio	**Etapas críticas**	**Referências**
Domínios críticos (Saúde, Finanças, Educação, Jurídico)	Coleta de dados; Treinamento; Monitoramento	(Chau et al., 2025; Krishnamurthy et al., 2025; Sinha et al., 2024)
Aplicações corporativas e industriais	Implantação; Monitoramento	(Bodor et al., 2025; Mahr et al., 2024; Shan et al., 2024)
Sistemas inteligentes e automação	Ajuste fino; Monitoramento; Implantação	(Krishnamurthy et al., 2025; Myakala et al., 2025)
Processamento de documentos (OCR e reconhecimento de caracteres)	Preparação de dados; Treinamento; Ajuste fino; Avaliação; Implantação	(Abdellatif et al., 2025; Gupta et al., 2024)
Engenharia de software	Avaliação	(Radenkovic et al., 2025)

Fonte: Elaborado pela autora (2026)

5.3 Limitações do estudo e possíveis vieses

Este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. Primeiramente, por se tratar de um *survey* da literatura, a análise depende diretamente dos estudos selecionados, podendo não abranger todas as abordagens existentes sobre LLMOps. Além disso, a área ainda é recente e em constante evolução, o que implica em uma possível limitação quanto à disponibilidade e maturidade dos trabalhos analisados.

No que se refere às ameaças à validade, destacam-se possíveis vieses no processo de seleção dos estudos, uma vez que, mesmo com a definição de critérios de inclusão e exclusão, a escolha das bases de dados e das estratégias de busca pode ter influenciado os resultados obtidos. Adicionalmente, a interpretação e categorização das etapas e interações do ciclo de vida de

LLMOps foram realizadas com base na análise dos autores deste trabalho, o que pode introduzir subjetividade na organização e na síntese das informações.

Em suma, ressalta-se que as análises realizadas consideram o contexto e os objetivos definidos neste estudo, podendo não refletir integralmente todas as variações possíveis de aplicação de LLMOps em diferentes domínios.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo analisar a literatura existente sobre LLMOps, com foco na compreensão das interações contínuas entre as etapas de seu ciclo de vida e na forma como essas etapas são aplicadas em diferentes domínios de atuação. A partir da análise dos estudos selecionados, foi possível identificar que o ciclo de vida do LLMOps não ocorre de maneira linear, sendo caracterizado por interações contínuas entre suas etapas, especialmente entre treinamento, avaliação, implantação e monitoramento. Essas interações evidenciam o caráter iterativo desse processo, no qual os resultados obtidos em cada fase influenciam diretamente as etapas subsequentes.

Além disso, observou-se que a relevância das etapas do LLMOps varia de acordo com o domínio de aplicação. Em contextos críticos, como saúde e finanças, destacam-se etapas relacionadas ao monitoramento e à garantia de confiabilidade dos modelos, enquanto, em ambientes empresariais e industriais, a etapa de implantação, especialmente no que se refere à integração e à escalabilidade, assume maior importância. Já em domínios específicos, como o setor imobiliário, a etapa de treinamento, especialmente por meio do ajuste fino, aliada ao monitoramento contínuo, mostra-se fundamental para lidar com a dinamicidade dos dados.

Dessa forma, os resultados obtidos permitem responder às questões de pesquisa propostas, evidenciando tanto a natureza interdependente das etapas do LLMOps quanto a influência do contexto de aplicação na definição de suas prioridades. Como contribuição, este estudo apresenta uma análise integrada dessas dimensões, preenchendo lacunas identificadas na literatura, especialmente no que se refere à interação entre etapas e à comparação entre diferentes domínios.

Por fim, destaca-se que, apesar das contribuições apresentadas, ainda existem oportunidades para investigações futuras, especialmente no que se refere à análise da aplicação de LLMOps em cenários reais, considerando aspectos como desafios de implantação, monitoramento em produção e adaptação contínua dos modelos a dados dinâmicos. Além disso, observa-se a necessidade de desenvolvimento de modelos mais padronizados que considerem explicitamente as interações contínuas entre as etapas do ciclo de vida, bem como a definição de diretrizes e boas práticas que auxiliem na integração dessas etapas em diferentes domínios de aplicação. Tais investigações podem ainda explorar a avaliação da eficiência dessas práticas em termos de desempenho, custo e confiabilidade. Esses avanços podem contribuir para o aprimoramento das práticas de LLMOps e para sua adoção mais eficiente em diferentes contextos.

REFERÊNCIAS

ABDELLATIF, O. H.; AYMAN, A.; HAMDI, A. Llmops-driven robotic process automation approach. In: **Proceedings of the IEEE International Conference on Intelligent Methods, Systems, and Applications (IMSA)**. [*S. l.*]: IEEE, 2025.

BODOR, A.; HNIDA, M.; DAOUDI, N. Integration of web scraping, fine-tuning, and

data enrichment in a continuous monitoring context via large language model operations. **International Journal of Electrical and Computer Engineering**, v. 15, n. 1, p. 1027–1037, 2025.

BOMMASANI, R. *et al.* On the opportunities and risks of foundation models. **arXiv**, 2021. Disponível em: https://arxiv.org/abs/2108.07258.

BROWN, T. *et al.* Language models are few-shot learners. **arXiv**, 2020. Disponível em: https://arxiv.org/abs/2005.14165.

CHAU, M.; XU, J. An is research agenda on large language models: Development, applications, and impacts on business and management. **ACM Transactions on Management Information Systems**, ACM, v. 16, n. 1, p. 1–11, fev. 2025.

CHEN, X.; LI, Y.; WANG, X. Design principles and guidelines for llm observability: Insights from developers. In: **Proceedings of the CHI Conference on Human Factors in Computing Systems**. [*S. l.*]: ACM, 2025.

CHERNIGOVSKAYA, M.; WALIA, D. S.; NEUMANN, K.; HARDT, A.; NAHHAS, A.; TUROWSKI, K. Towards a standardized business process model for llmops. In: . [*S. l.*: *s. n.*], 2025.

DIAZ-DE-ARCAYA, J.; LóPEZ-DE-ARMENTIA, J.; MIñóN, R.; OJANGUREN, I. L.; TORRE-BASTIDA, A. I. Large language model operations (llmops): Definition, challenges, and lifecycle management. In: **Proceedings of the 2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)**. [*S. l.*]: IEEE, 2024. p. 1–6.

GUPTA, U.; KUMAR, A.; GUPTA, A.; RAJ, G.; AGRAWAL, A. P. Advances in handwritten character recognition: A comparison of ocr and large language model-based approaches. In: **Proceedings of the IEEE International Conference on Emerging Innovation (EmergIN)**. [*S. l.*]: IEEE, 2024.

IBM. **What are large language models?** 2023. IBM website. Disponível em: https://www.ibm.com/topics/large-language-models.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. v. 2, 01 2007.

KREUZBERGER, D.; KüHL, N.; HIRSCHL, S. Machine learning operations (mlops): Overview, definition, and architecture. **IEEE Access**, 2023.

KRISHNAMURTHY, A.; NEELANATH, N. Establishing a robust llmops framework for intelligent automation: Strategies and best practices. **arXiv**, 2025. Disponível em: https://ieeexplore.ieee.org/abstract/document/10961869.

LIU, S.; FARKIANI, S. N.; CROWLEY, J. L. A survey on large language models for network operations and management: applications, techniques, and opportunities. **arXiv**, 2024.

MAHR, F.; SCHMIDT, K.; ANGELI, G.; FRANKE, J.; SINDEL, T. A reference architecture for deploying large language model applications in industrial environments. In: **Proceedings of the IEEE 30th International Symposium for Design and Technology in Electronic Packaging (SIITME)**. Sibiu, Romania: IEEE, 2024. p. 19–23.

MIAO, X.; JIA, Z.; CUI, B. Demystifying data management for large language models. In: **Companion of the 2024 International Conference on Management of Data**. New York, NY, USA: Association for Computing Machinery, 2024. (SIGMOD '24), p. 547–555. ISBN 9798400704222. Disponível em: https://doi.org/10.1145/3626246.3654683.

Microsoft. **LLMOps - Operational management of LLMs**. 2025. Acesso em: 28 abr. 2026. Disponível em: https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/mlops-in-openai/.

MYAKALA, P. K.; KAMATALA, S.; NAAYINI, P. Edge ai and federated llmops for latency-critical iot systems in smart infrastructure. In: **Proceedings of the International Conference on Intelligent Technologies (CONIT)**. [*S. l.*]: IEEE, 2025.

PAHUNE, S.; AKHTAR, Z. Transitioning from mlops to llmops: Navigating the unique challenges of large language models. **Information**, v. 16, n. 2, p. 87, 2025. Disponível em: https://www.mdpi.com/2078-2489/16/2/87.

PATIL, R.; GUDIVADA, V. A review of current trends, techniques, and challenges in large language models (llms). **Applied Sciences**, v. 14, n. 5, p. 2074, 2024. Disponível em: https://doi.org/10.3390/app14052074.

POLYAKOVSKA, N. Towards a practical framework for llmops: Understanding and building operational excellence for large language models. **Computer-Integrated Technologies: Education, Science, Production**, Lutsk, n. 58, 2025.

RADENKOVIC, B.; PROKHOROV, S.; RADENKOVIC, M.; LABUS, A. Application of large language models in software development: Review of the current state and development perspectives. In: **Proceedings of the IEEE Conference on Emerging Technologies (EnT)**. [*S. l.*]: IEEE, 2025.

RAVINDRAN, S. K. Unified threat detection and mitigation framework (UTDMF): Combating prompt injection, deception, and bias in enterprise-scale transformers. **arXiv**, 2025. Acesso em: 28 abr. 2026. Disponível em: https://ar5iv.labs.arxiv.org/html/2510.04528.

RAZA, S.; SAPKOTA, R.; KARKEE, M.; EMMANOUILIDIS, C. Trust, risk, and security in agentic AI: A short survey. In: **Proceedings of the Neural Information Processing Systems (NeurIPS)**. San Diego, CA, USA: [*S. n.*], 2025. Acesso em: 28 abr. 2026. Disponível em: https://nips.cc/virtual/2025/loc/san-diego/137151.

RICHARDSON, W. S.; WILSON, M. C.; NISHIKAWA, J.; HAYWARD, R. S. The well-built clinical question: a key to evidence-based decisions. **ACP journal club**, v. 123 3, p. A12–3, 1995. Disponível em: https://api.semanticscholar.org/CorpusID:32595527.

SHAN, R.; SHAN, T. Enterprise llmops: Advancing large language models operations practice. In: **Proceedings of the IEEE Cloud Summit**. [*S. l.*]: IEEE, 2024. p. 143–148.

SHI, C.; LIANG, P.; WU, Y.; ZHAN, T.; JIN, Z. Maximizing user experience with llmops-driven personalized recommendation systems. **Information**, 2024. Disponível em: https://ace.ewapub.com/article/view/12388.pdf.

SINHA, M.; MENON, S.; SAGAR, R. Llmops: Definitions, framework and best practices. In: **Proceedings of the International Conference on Electrical, Computer and Energy**

Technologies (ICECET). IEEE, 2024. Disponível em: https://ieeexplore.ieee.org/document/10698359.

THOMPSON, A. Leveraging devops and mlops to enhance feedback loops in machine learning model development. **African Journal of Artificial Intelligence and Sustainable Development**, v. 4, n. 2, 2024.

Tredence. **LLMOps Lifecycle: A Comprehensive Guide**. 2025. Acesso em: 26 mar. 2026. Disponível em: https//www.tredence.com/blog/llmops-lifecycle.

VASWANI, A. *et al.* Attention is all you need. **arXiv**, 2017. Disponível em: https://arxiv.org/abs/1706.03762.

XU, X.; WEYTJENS, H.; ZHANG, D.; LU, Q.; WEBER, I.; ZHU, L. RAGOps: Operating and managing retrieval-augmented generation pipelines. **arXiv**, abs/2506.03401, 2025. Acesso em: 28 abr. 2026. Disponível em: https://arxiv.org/pdf/2506.03401.

ÖZER, F. C. **Systematic Technical Survey on LLMOps: Lifecycle, Tools, Challenges, and Emerging Practices**. 2025. Acesso em: 25 mar. 2026. Disponível em: https://erepo.uef.fi/items/66c019d1-a67a-4155-928c-bcc1357c4e77.