RAG e Gamificação no Ensino de MLOps: Avaliação da Precisão e Relevância de um Chatbot Educacional por "LLM as Judge"

Luiz Gomes Ribeiro Neto
Globo
Rio de Janeiro, Brasil
luiz.neto@g.globo

Edison Fabián Caballero Pérez
Globo
Rio de Janeiro, Brasil
fabian.caballero@g.globo

Guilherme Vinícius Lima dos Anjos
Globo
Rio de Janeiro, Brasil
guilherme.vinicius@g.globo

Jacson Rodrigues Barbosa
Centro de Competência Embrapii em Tecnologias Imersivas (AKCIT)/Instituto de Informática - UFG
Goiás, Brasil
jacson_rodrigues@ufg.br

Resumo

O ensino de Machine Learning Operations (MLOps) enfrenta desafios significativos devido à sua natureza abstrata, o que frequentemente resulta na desmotivação de estudantes e profissionais recém-ingressados no mercado. Diante desse cenário, esta pesquisa teve como objetivo elaborar e validar um sistema de tutoria adaptativa que utiliza mecânicas de gamificação, como narrativas e recompensas, para promover um aprendizado mais imersivo e engajador via chatbot. A tecnologia desenvolvida fundamenta-se na arquitetura Retrieval-Augmented Generation (RAG), projetada especificamente para mitigar as alucinações comuns em modelos de linguagem e garantir que as respostas fornecidas sejam tecnicamente precisas e fiéis a uma base de conhecimento validada. O desenvolvimento seguiu um rigoroso percurso metodológico de métodos mistos, organizado em cinco fases sequenciais: fundamentação teórica, construção de um Ground Truth, implementação do protótipo funcional, otimização técnica de parâmetros e coleta de feedbacks. A validação da solução foi realizada através da metodologia inovadora "LLM as Judge", em conjunto com a análise do feedback qualitativo dos usuários finais. O estudo conclui que a tecnologia facilita a aquisição de conhecimentos técnicos em MLOps e estabelece diretrizes para a criação de sistemas educacionais mais robustos e inovadores.

Palavras-chave

Machine Learning Operations, Inteligência Artificial, Inteligência Artificial Generativa, Retrieval Augmented Generation

1 Introdução

O Machine Learning Operations (MLOps) consolidou-se como um paradigma essencial para gerenciar o ciclo de vida de modelos de aprendizado de máquina em escala, unindo práticas de Machine Learning, DevOps e Engenharia de Dados. O MLOps atua como uma cultura que visa reduzir a dívida técnica inerente aos sistemas de IA [2]. No entanto, o setor enfrenta um hiato crítico, frequentemente referido como o "vale da morte" da IA, onde até 88% das iniciativas corporativas não conseguem ultrapassar a prototipagem [9]. Essa alta taxa de falha ocorre devido à desconexão entre o desenvolvimento do modelo em laboratório e os rigorosos requisitos operacionais de produção [6].

No contexto educacional, essa complexidade estrutural gera barreiras pedagógicas severas. O ensino de MLOps exige uma abordagem multidisciplinar que sobrecarrega a carga cognitiva dos alunos, demandando competências simultâneas em estatística, software e infraestrutura [4]. Essa curva de aprendizado íngreme frequentemente resulta na desmotivação e na evasão de estudantes que lutam para visualizar a aplicação prática dos conceitos sem o apoio de ambientes que simulem cenários reais [7].

Diante desse cenário, os Large Language Models (LLMs) oferecem um potencial transformador ao permitir uma tutoria instantânea e personalizada. Contudo, a propensão inerente a "alucinações" - a geração de conteúdos linguisticamente fluidos, mas factualmente incorretos - representa um risco inaceitável no ensino de engenharia [1]. Para mitigar esse problema, este trabalho propõe um sistema de tutoria adaptativa fundamentado na arquitetura Retrieval-Augmented Generation (RAG), que "ancora" a geração de texto em uma base de conhecimento validada para garantir precisão técnica [3]. Complementarmente, para combater a evasão, integra-se uma camada de gamificação baseada na Teoria da Autodeterminação [5].

O objetivo geral deste estudo é elaborar e validar este chatbot educacional de MLOps. Dada a dificuldade de escalar a avaliação manual de IAs conversacionais, a pesquisa adota a metodologia de vanguarda "LLM as Judge" [10] para verificar a qualidade das respostas de forma objetiva, estabelecendo um framework de ensino técnico que seja simultaneamente robusto, seguro e pedagogicamente inovador.

2 Descrição da Tecnologia

A tecnologia desenvolvida é um chatbot educacional imersivo projetado para transpor a barreira da abstração no ensino de MLOps, integrando inteligência artificial generativa, práticas pedagógicas e elementos de jogos.

A base técnica utiliza a arquitetura RAG, ancorando as respostas em uma base de conhecimento externa curada por especialistas humanos (como artigos e documentação técnica), priorizando fontes oficiais e evitando redundâncias conflitantes entre documentos para reduzir o risco de recuperação de trechos contraditórios. O sistema utiliza modelos de vector embeddings para busca semântica, recuperando o contexto pertinente, o que reduz alucinações e permite a citação de fontes verificáveis para garantir a transparência.

A camada de gamificação, fundamentada na Teoria da Autodeterminação, incorpora um sistema de progressão baseado na dificuldade dos tópicos (iniciante, intermediário e avançado), para satisfazer necessidades de competência e autonomia, combatendo a desmotivação. O agente busca atuar na Zona de Desenvolvimento Proximal (ZDP), que é definida como a distância entre o nível de desenvolvimento atual (o que o aluno consegue resolver sozinho) e o nível de desenvolvimento potencial (o que ele consegue resolver com orientação) [8].

A otimização técnica envolve o uso de frameworks como LangChain, Hugging Face, ChromaDB e Google ADK, com calibração contínua de hiperparâmetros da base vetorial (como tamanho de chunks de texto e overlap).

2.1 Procedimentos Metodológicos

O desenvolvimento seguiu uma abordagem de métodos mistos (Mixed Methods Research), estruturada em cinco fases sequenciais e iterativas:

(1) `Fundamentação e Curadoria`: Pesquisa bibliográfica para compor a base de conhecimento focada nas melhores práticas de MLOps.
(2) `Construção do Ground Truth`: Elaboração de um padrão ouro de 36 perguntas e respostas separadas em três níveis de dificuldade e revisado por especialistas, essencial para calibrar a avaliação automática.
(3) `Desenvolvimento do Protótipo`: Implementação da arquitetura RAG e integração da camada conversacional gamificada utilizando Python, LangChain e Google ADK.
(4) `Otimização Paramétrica`: Realização de testes para refinar hiperparâmetros críticos, como tamanho de fragmentos (chunks).
(5) `Avaliação Híbrida`: Avaliação automatizada quantitativa utilizando a metodologia "LLM as a Judge", na qual um comitê de três modelos (Llama 3.1, Llama 3.2 e GPT-OSS) atribuiu notas de 0 a 1 às respostas geradas frente ao conjunto de Ground Truth revisado por especialistas, segundo quatro métricas do framework DeepEval (Context Precision, Context Relevancy, Answer Relevancy e Faithfulness) e avaliação qualitativa por meio de questionários (System Usability Scale) aplicados a usuários com diferentes níveis de senioridade.

3 Testes e Resultados

A **Fase 1** primeiramente avaliou o desempenho da RAG frente a oito configurações de *chunk size* e *chunk overlap*, utilizando o GPT-OSS como modelo generativo (Tabela 1). Essa análise paramétrica orientou a escolha da configuração chunk size = 1000 e chunk overlap = 200, adotada nos testes subsequentes por apresentar o melhor equilíbrio entre as quatro métricas avaliadas.

Após a calibragem, seguimos para os modelos juízes avaliadores, onde observamos comportamentos distintos:

- `Relevância da Resposta (Answer Relevancy)`: Quando avaliado pelo GPT-OSS, o modelo gerador Llama 3.1 obteve pontuações de excelência sistêmica, com notas situadas predominantemente na faixa máxima entre 0.9 e 1.0. Isso atesta que as respostas são diretas e livres de tangentes, um requisito vital para desenvolvedores que buscam soluções rápidas durante o coding.
- `Fidelidade Factual (Faithfulness)`: O Llama 3.2 destacou-se em sua aderência ao contexto fornecido. A evidência numérica sugere que este modelo mitiga a ocorrência de alucinações, tornando-se seguro para aplicações de engenharia onde um comando incorreto pode gerar falhas críticas em produção.
- `Relevância de Contexto (Context Relevancy)`: O Llama 3.1 atuou como um "juiz severo", atribuindo notas mais baixas e penalizando excesso de ruídos na recuperação de informação do banco de dados vetorial, forçando a equipe a calibrar os hiperparâmetros de chunking para um nível de precisão mais alto.

Diante desses resultados, o GPT-OSS foi selecionado como modelo generativo do agente na Fase 2 por apresentar o melhor equilíbrio entre as quatro métricas avaliadas, com desempenho levemente superior e mais estável do que o Deepseek-r1 em precisão e relevância de contexto, ainda que sem liderar isoladamente nenhuma métrica, evidenciando que a combinação de múltiplos juízes automáticos ajuda a mitigar o viés individual de cada avaliador.

A **Fase 2** (avaliação qualitativa com usuários) demonstrou uma adaptabilidade bem-sucedida aos diferentes níveis de conhecimento. Para obter provas de evidência claras sobre o impacto da ferramenta no cotidiano da equipe de desenvolvimento de software, foram aplicados questionários fundamentados na System Usability Scale (SUS), nos quais os usuários avaliaram sete dimensões da ferramenta em uma escala numérica de 0 a 5 (onde 0 é o impacto mais negativo e 5 o mais positivo).

Os resultados, consolidados no gráfico de radar representado na Figura 1 da avaliação qualitativa, forneceram números concretos do impacto da solução em diferentes frentes da equipe de desenvolvimento:

- `Impacto Operacional na Equipe de Infraestrutura (DevOps)`: O perfil responsável pela sustentação de software (Usuário DevOps) atribuiu pontuação máxima (nota 5) nas dimensões de "Personalização", "Impacto" e "Confiabilidade".
- `Acurácia Técnica para a Equipe de Modelagem (DS Júnior)`: O Cientista de Dados Júnior, que já possui embasamento para julgar tecnicamente as respostas, atribuiu a nota máxima (5) à dimensão de "Qualidade".
- `Retenção e Amparo de Novos Desenvolvedores`: Profissionais recém-integrados (Estagiários) às equipes relataram picos de avaliação com pontuações máximas (nota 5) nas dimensões de "Engajamento" e "Suporte".

4 Discussão

A análise corrobora que a integração entre o rigor técnico da RAG e as estratégias pedagógicas de gamificação é viável e necessária. A avaliação técnica aponta que o Llama 3.2 é a opção superior para garantir fidelidade factual (requisito crítico no ensino de engenharia), enquanto o Llama 3.1 garante maior rigor na relevância semântica da geração.

A avaliação qualitativa via questionários baseados na System Usability Scale (SUS) forneceu indicadores quantitativos consistentes

Tabela 1: Pontuação média por métrica para cada configuração de chunk size e overlap (GPT-OSS).

Configuração Chunk Size e Overlap	Context Precision	Context Relevancy	Answer Relevancy	Faithfulness
Chunk Size=50 / Overlap=5	média=0.46 amostras=15	média=0.22 amostras=15	média=0.77 amostras=14	média=0.83 amostras=13
Chunk Size=100 / Overlap=10	média=0.69 amostras=34	média=0.44 amostras=35	média=0.68 amostras=34	média=0.69 amostras=34
Chunk Size=250 / Overlap=25	média=0.74 amostras=32	média=0.49 amostras=31	média=0.69 amostras=33	média=0.73 amostras=34
Chunk Size=500 / Overlap=50	média=0.67 amostras=33	média=0.48 amostras=27	média=0.66 amostras=35	média=0.74 amostras=34
Chunk Size=750 / Overlap=75	média=0.70 amostras=30	média=0.44 amostras=22	média=0.69 amostras=34	média=0.76 amostras=34
Chunk Size=1000 / Overlap=100	média=0.80 amostras=33	média=0.11 amostras=12	média=0.60 amostras=31	média=0.76 amostras=25
Chunk Size=2000 / Overlap=200	média=0.62 amostras=31	média=0.30 amostras=22	média=0.69 amostras=34	média=0.79 amostras=29
Chunk Size=1000 / Overlap=200	média=0.68 amostras=32	média=0.43 amostras=34	média=0.73 amostras=35	média=0.78 amostras=35

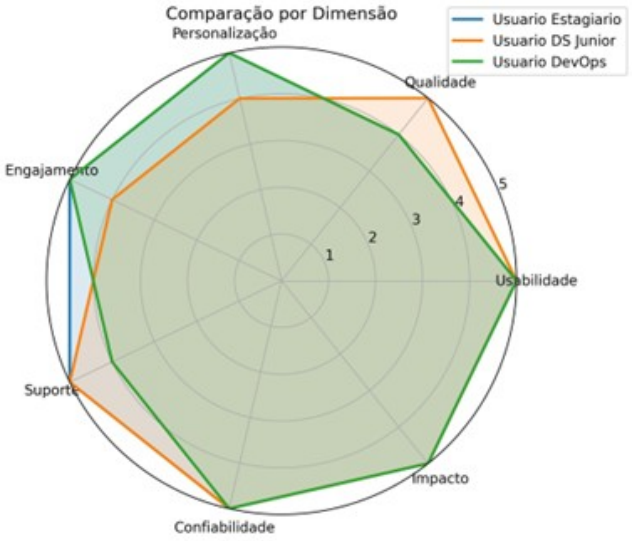

Figura 1: Comparação das dimensões avaliadas qualitativamente com os usuários.

de que o sistema possui um "sucesso híbrido", resolvendo dores específicas de diferentes papéis dentro do ciclo de desenvolvimento de software:

- Apropriação Tecnológica para DevOps: A transição de um modelo de IA do laboratório para a produção (o "vale da morte" mencionado anteriormente) exige que a equipe de infraestrutura compreenda a fundo a solução para mantê-la. O perfil DevOps atribuiu a pontuação máxima (5/5) nas dimensões de "Personalização", "Impacto" e "Confiabilidade". Essa evidência sugere que o chatbot não apenas repassa teoria, mas pode capacitar o profissonal a realizar a sustentação do software com independência e velocidade.
- Alinhamento de Regras de Negócio para o DS Júnior: Projetos de IA sofrem quando a equipe de modelagem não consegue traduzir conceitos acadêmicos para necessidades reais da aplicação final. O perfil DS Júnior atribuiu nota máxima (5/5) em "Qualidade". Essa avaliação indica que o chatbot oferece o nível de profundidade adequado ao exigido para que o cientista de dados aprenda os conceitos e eventualmente tome decisões arquiteturais coerentes com o negócio.
- Onboarding e Redução de Atrito para Estagiários: O turnover e a entrada contínua de novos membros nas equipes podem gerar instabilidade. Para os estagiários, o chatbot obteve pontuações máximas (5/5) em "Engajamento" e "Suporte". Esse resultado indica que as mecânicas de gamificação e o design conversacional podem funcionar como uma ponte de onboarding, reduzindo a fricção cognitiva, amparando os iniciantes diante da complexidade do código e diminuindo a sobrecarga dos desenvolvedores seniores que atuariam como tutores.

Em suma, os resultados evidenciam que o sistema fundamentado em RAG e Gamificação vai muito além da sala de aula: ele atua diretamente no gargalo de apropriação de conhecimento das equipes técnicas, facilitando a transferência tecnológica e garantindo que o ciclo de vida do modelo (MLOps) seja compreendido, implementado e sustentado com eficácia pelas equipes de software.

É importante situar esses resultados dentro de suas limitações. As sete dimensões avaliadas com os usuários capturam principalmente qualidade percebida, confiança e engajamento, mas não constituem uma medida direta de ganho de aprendizado - uma limitação que trabalhos futuros podem endereçar com avaliações de conhecimento antes e depois do uso do agente. Da mesma forma, os resultados não isolam o efeito específico do RAG e da gamificação frente a uma alternativa mais simples, como uma LLM genérica respondendo diretamente a partir dos mesmos documentos; essa comparação ajudaria a quantificar o ganho real das escolhas arquiteturais adotadas. Por fim, a interface é hoje restrita ao texto conversacional, o que pode limitar o ensino de conceitos visuais e sequenciais próprios de pipelines de MLOps, e também alguns casos de acessibilidade. Também vale mencionar que a escolha dos modelos foi sujeita à disponibilidade de recursos computacionais presentes no momento.

5 Considerações Finais

A avaliação do chatbot educacional validou o que o estudo define como um "sucesso híbrido", alcançado ao integrar perfeitamente a segurança técnica da inteligência artificial à utilidade prática e pedagógica.

A validação técnica automatizada (Fase 1) comprovou a precisão da arquitetura RAG através da metodologia "LLM as a Judge". Os testes mostraram que é possível mitigar de forma drástica o risco de alucinações ao utilizar o modelo Llama 3.2 (que garantiu excelência em fidelidade factual) em conjunto com o Llama 3.1 (que agiu como um filtro rigoroso para manter a relevância do contexto recuperado).

Essa robustez comprovada nos bastidores refletiu-se diretamente na percepção da equipe técnica de software (Fase 2). A ferramenta demonstrou uma adaptabilidade notável aos diferentes níveis de conhecimento dos usuários: para os iniciantes (Estagiários), os elementos de gamificação funcionaram perfeitamente como um mecanismo de engajamento e suporte, reduzindo a frustração do aprendizado inicial. Simultaneamente, a exatidão validada na primeira fase garantiu que profissionais seniores (DevOps e Cientistas de Dados Júnior) atribuíssem notas máximas para a qualidade, personalização e confiabilidade da ferramenta ao lidar com desafios práticos e reais de modelagem e infraestrutura.

Em suma, a correlação entre as duas fases conclui que o sistema não apenas gera respostas tecnicamente corretas e seguras (Fase 1), mas consegue entregá-las com a abordagem e o amparo adequados para alavancar a produtividade e o aprendizado de toda a equipe de desenvolvimento (Fase 2).

Como próximos passos, recomenda-se a expansão da base de dados de conhecimento para abranger tópicos intermediários e avançados (como governança em IA), a implementação de desafios práticos (hands-on) e narrativas mais elaboradas ao longo das interações, incluindo a evolução para o modo socrático, hoje limitado à progressão por dificuldade, para perguntas graduais que estimulem a reflexão do aluno antes de fornecer a resposta direta. Adicionalmente, sugere-se a criação de um mecanismo de avaliação, como uma prova final onde o aluno só passa para a próxima seção se obter uma nota mínima, o que também permitiria medir de forma mais direta o ganho de aprendizado proporcionado pela ferramenta. Para bases de conhecimento maiores, técnicas como graphrag também podem ser exploradas como alternativa ao RAG tradicional.

Também são necessários testes adicionais que ataquem diretamente as limitações dos resultados atuais: avaliações de conhecimento antes e depois do uso do agente para mensurar o ganho de aprendizado de forma direta, testes comparativos entre o agente e uma LLM genérica respondendo a partir dos mesmos documentos para isolar o efeito específico do RAG e da gamificação, e estudos com recursos visuais complementares ao chat para verificar seu impacto no ensino de conceitos sequenciais e espaciais de MLOps, além de melhorar a acessibilidade da ferramenta.

Disponibilidade de Artefatos

Os artefatos produzidos neste trabalho (protótipo do agente e conjunto de Ground Truth) não estão publicamente disponíveis no momento da publicação, uma vez que o protótipo, encontra-se em avaliação para eventual produtização interna.

Referências

[1] Ziwei Ji et al. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[2] Dominik Kreuzberger et al. 2023. Machine learning operations (MLOps): overview, definition, and architecture. *IEEE Access* 11 (2023), 31866–31879.

[3] Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. Curran Associates, Inc., 9459–9474.

[4] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. 2022. Challenges in deploying machine learning: a survey of case studies. *Comput. Surveys* 55, 6 (2022), 1–29. doi:10.1145/3533378

[5] Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55, 1 (2000), 68–78.

[6] David Sculley et al. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 28. Montreal, 2503–2511.

[7] Jakub Swacha and Marcin Gracel. 2025. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Applied Sciences* 15, 8 (2025), 4234. doi:10.3390/app15084234

[8] Oren Vainas, Oren Bar-Ilan, Yaniv Ben-David, Ran Gilad-Bachrach, Gal Lukin, Michael Ronen, Ron Shillo, and Doron Sitton. 2019. E-Gotsky: Sequencing Content Using the Zone of Proximal Development. *arXiv preprint arXiv:1904.12268* (2019). arXiv:1904.12268 [cs.CY] https://arxiv.org/abs/1904.12268

[9] VB STAFF. 2019. Why do 87% of data science projects never make it into production? VentureBeat. https://venturebeat.com/technology/why-do-87-of-data-science-projects-never-make-it-intoproduction Acesso em: 14 fev. 2026.

[10] Lianmin Zheng et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*. New Orleans, LA, USA.