Teaching Assistant Based on a Brazilian Large Language Model

Elton S. Siqueira
eltonss@ufpa.br
Federal University of Pará - Faculty of Information Systems
Cametá, Pará, Brazil

Carlos S. Portela
csp@ufpa.br
Federal University of Pará - Faculty of Information Systems
Cametá, Pará, Brazil

Augusto N. Moraes
augusto.nonato.moraes@gmail.com
Federal University of Pará - Faculty of Information Systems
Cametá, Pará, Brazil

Abstract

Context: The use of AI Teaching Assistants (AI-TAs) in educational contexts aims to support student learning in complex subjects, where there are challenges in assimilating interdisciplinary concepts and a need for quick and accurate responses. **Problem:** Students face barriers when studying subjects that require prior knowledge in multiple areas, leading to comprehension difficulties that negatively affect their academic progress. **Proposed Solution:** An AI-TA based on the Brazilian language model Sabiá 2.0 is proposed, incorporating mechanisms for Quick Verification, Expansion, and Error Correction to enhance the question-and-answer system. **IS Theory:** This work is grounded in Social Learning Theory, which explores how individuals acquire knowledge in technology-mediated interaction contexts. This theory is applicable to the use of AI-TAs, which act as learning mediators, promoting knowledge acquisition and facilitating understanding of complex concepts. **Method:** The methodology is Design Science Research, with a descriptive approach and quantitative analysis, including proof of concept and practical evaluation with users. The system was implemented in an educational environment and tested for performance and student satisfaction. **Summary of Results:** The system showed a high satisfaction rate regarding response quality; however, it revealed limitations in response time for model-based queries, suggesting improvements in API optimization. **Contributions and Impact in IS Field:** This study contributes to the IS field by demonstrating how AI models can enhance teaching and student support, proposing a robust and adaptable solution for learning environments.

CCS Concepts

• **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**; • **Discourse, dialogue and pragmatics**;

Keywords

Educational Support, Teaching Assistant, LLM, Question-Answering System.

1 Introdução

A expansão das tecnologias de Inteligência Artificial (IA) no contexto educacional está permitindo a produção de uma grande quantidade de pesquisas sobre o uso de Modelo de Linguagem de Grande Escala (*Large Language Model* - LLMs) como assistentes de ensino, explorando como esses sistemas permitem apoiar os profissionais da educação em várias tarefas [7]. Por exemplo, LLMs como o ChatGPT têm sido integrados em salas de aula para ajudar na geração de perguntas, correção automática e orientação pedagógica. Esses sistemas podem produzir perguntas de compreensão de leitura, problemas matemáticos e dicas para orientar os alunos sem fornecer soluções diretas. Embora sejam eficazes em reduzir a carga de trabalho dos educadores, automatizando tarefas rotineiras, a orientação oferecida pelos LLMs frequentemente necessita de supervisão dos instrutores humanos para garantir precisão e adequação aos diferentes níveis de aprendizado [13].

Um desafio importante ao usar LLMs é evitar a dependência excessiva, uma vez que alunos e professores podem simplesmente aceitar o conteúdo gerado pela IA sem avaliá-lo criticamente. Isso pode ser mitigado combinando a expertise humana com o suporte da IA para melhorar o aprendizado, ao mesmo tempo que mantém habilidades de pensamento crítico e resolução de problemas.

As pesquisas também destacam que os LLMs podem gerar questões e/ou respostas alinhadas com objetivos de aprendizagem específicos, oferecendo aos professores e alunos uma ferramenta poderosa para criar e obter materiais educacionais. No entanto, questões relacionadas a qualidade do conteúdo e veracidade das informações podem sofrer algumas distorções durante o processo de geração textual, comprometendo o processo de aprendizagem [6].

Outro problema deparado por muitos acadêmicos é a deficiência de conhecimento básico em campos interdisciplinares, que ocorre em cursos que relacionam vários conceitos de diversos campos do conhecimento, como computação, matemática e engenharia. Conforme Ahmed, Jeon e Piccialli [1] ressaltam que essa lacuna pode impactar no sucesso acadêmico, principalmente em disciplinas interdisciplinares (por exemplo, Algoritmos), e que o uso da IA na educação tem capacidade de auxiliar no processo de aprendizagem, oferecendo suporte personalizado e adaptado as necessidades individuais dos acadêmicos. Gong [4] observou que tecnologias emergentes, como realidade virtual e aumentada, podem promover e facilitar a integração do conhecimento em diversos campos, auxiliando a preencher essas lacunas e facilitar o aprendizado aplicado.

O objetivo geral deste trabalho é apresentar uma solução no contexto educacional utilizando um Assistente de Ensino baseado em IA que utiliza LLM brasileiro, que se destina ao suporte de alunos durante o processo de aprendizagem por meio de um sistema de Perguntas e Respostas (*Questions & Answers - Q&A*). O trabalho visa minimizar os impactos relacionados a deficiência de fundamentos essenciais enfrentados por estudantes ao cursarem disciplinas que exigem conhecimento prévio. Para alcançar o objetivo geral deste trabalho, são definidos os seguintes objetivos específicos:

- Implementar um Mecanismo de Verificação Rápida (MVR): Otimizar o tempo de resposta do assistente ao identificar perguntas similares, promovendo respostas imediatas para questões já conhecidas.
- Desenvolver um mecanismo de expansão de perguntas: Expandir o banco de questões utilizando a API Maritaca para

gerar variações semânticas, garantindo flexibilidade e adaptabilidade do sistema às diversas formulações de perguntas dos usuários.
- Implementar um Mecanismo de Correção de Erros (MCE): Melhorar a precisão das respostas com base no *feedback* dos usuários, possibilitando um aprimoramento contínuo e maior confiabilidade no sistema de perguntas e respostas.
- Avaliar a efetividade e satisfação dos usuários: Medir a satisfação dos estudantes com a qualidade e o tempo das respostas, utilizando métricas quantitativas e qualitativas para ajustar o desempenho do sistema.

A arquitetura do sistema proposto neste estudo inclui três módulos: I) Cliente; II) Servidor e dataset; III) API Maritaca. Os usuários podem usar o módulo cliente em seus dispositivos móveis ou computadores para realizar as perguntas, e o módulo servidor responde as perguntas dos usuários. Além disso, este trabalho propõe um mecanismo de expansão, em que o módulo servidor realiza uma chamada via API Maritaca [10, 2] para combinar a tecnologia de geração de texto para expandir o banco de questões (dataset). Além da expansão de questões, foi incluído um mecanismo de correção de erros para que o sistema melhore suas respostas com base no *feedback* do usuário, aumentando a qualidade e relevância das respostas ao longo do tempo. Um outro ponto relevante deste trabalho é a implementação de um mecanismo de verificação rápida que possui a responsabilidade de verificar se existe alguma pergunta similar no banco de questões. Esta abordagem não somente melhora a precisão do sistema de Q&A [8], mas também sua capacidade de lidar com a diversidade de perguntas dos usuários.

As principais contribuições deste artigo são:

- No campo do ensino, oferece suporte contínuo aos estudantes com respostas imediatas e confiáveis, consequentemente aprimorando o processo de aprendizagem.
- No campo científico, a contribuição está relacionada ao fomento de novas pesquisas relacionadas à IA aplicada à educação a partir do desenvolvimento de um sistema de Q&A que utiliza modelo de linguagem brasileiro. Ainda, propõe a expansão do banco de perguntas através da API Maritaca, gerando diferentes formulações para a mesma questão. Essa abordagem visa melhorar a precisão e a robustez do sistema de Q&A no servidor.

Este artigo está organizado da seguinte maneira: A Seção 2 consiste na revisão da literatura, que apresenta uma visão geral das pesquisas anteriores relacionadas ao tema. Em seguida, a Seção 3 descreve os elementos que compõem o sistema. Já a Seção 4 descreve a metodologia, detalhando os métodos e técnicas utilizados para a construção e validação do sistema. Adicionalmente, a Seção 5 apresenta os resultados obtidos com a aplicação do sistema e discute suas implicações. Por fim, a Seção 6 finaliza com a conclusão, que sumariza os principais achados deste estudo e sugere direções para pesquisas futuras.

2 Revisão da Literatura

As pesquisas sobre o uso de assistentes virtuais de ensino baseados em inteligência artificial têm avançado, abrindo novas perspectivas para a melhoria da experiência de aprendizado dos estudantes. Ao explorar diversas abordagens e tecnologias, essas pesquisas buscam não somente otimizar o processo de ensino, mas promover uma integração mais eficiente do conhecimento em ambientes interdisciplinares. Nesse contexto, o trabalho feito por [3] propõe uma solução baseada em um Assistente Virtual de Ensino que combina um sistema de Perguntas e Respostas com o ChatGPT. Essa integração permite que os alunos recebam respostas de maneira correta, criando uma ferramenta eficiente para o esclarecimento de dúvidas e oferecendo suporte adicional ao processo de aprendizado, de forma a melhorar a experiência educacional em tempo real. Porém, o referido trabalho não considera as percepções humanas para validação das respostas e o processo de correção ocorre usando uma rede neural.

Apesar das várias oportunidades relacionadas aos Modelos de Linguagem (LLMs) aplicados na educação, também existem desafios relacionados a codificação que precisam ser enfrentados. Como observado por [7], ferramentas como o ChatGPT têm um grande potencial para auxiliar tanto professores quanto alunos. Elas podem desempenhar o papel de assistentes na criação de conteúdos, além de oferecer *feedback* detalhado, ou apoiando em tarefas mais complexas, como a personalização do suporte pedagógico. Essas capacidades não só ampliam o alcance do ensino, mas também trazem uma nova camada de interação ao processo educativo. No entanto, o estudo ressalta um ponto fundamental: a necessidade de uma supervisão humana no uso da IA. Sem essa intervenção humana, corre-se o risco das respostas automáticas dos modelos de IA gerarem certa confusão ou distorção para usuário, o que pode prejudicar o desenvolvimento da criatividade e do pensamento crítico, habilidades essenciais para a formação de qualquer estudante.

Outro ponto destacado na literatura é a aplicação de LLMs em contextos interativos e de aprendizagem baseada em jogos. De acordo com Huber et al. [6] que exploram como a aprendizagem lúdica e gamificada pode se beneficiar com uso dos LLMs, promovendo um ambiente em que os estudantes são incentivados a praticar habilidades e a desenvolver competências específicas. A referida abordagem ajuda a reduzir o risco do aluno ficar dependente ou uso excessivo da tecnologia, em vez disso, ele seria capaz de usar os LLMs como um mera ferramenta de auxílio, ao mesmo tempo em que se envolve na prática do raciocínio crítico e resolução de problemas. Já o trabalho feito por Hicke et al. [5] examina o uso de LLMs para responder a perguntas em plataformas de ensino online, como *Piazza*, propondo a integração de técnicas como Geração Aumentada por Recuperação (RAG) e Otimização de Preferências Diretas (DPO) para melhorar a qualidade das respostas. Este estudo demonstrou um aumento significativo na precisão das respostas geradas e sugere que a combinação de LLMs com técnicas de ajuste fino pode proporcionar um suporte mais eficaz para os alunos.

No mesmo sentido, Moreira et al. [9] pesquisaram a utilização de Sistemas Tutores Inteligentes (STIs) associados a jogos educacionais digitais, com a finalidade de engajar e capacitar os estudantes em identificar *fake news*. O resultado deste trabalho está diretamente ligado a personalização dos STIs conforme as necessidades particulares dos discentes, desta forma tornando o processo de aprendizagem eficaz. O estudo observou uma melhoria no desempenho dos alunos que utilizaram o sistema com o tutor inteligente em comparação aos que jogaram sem o tutor, demonstrando o valor da personalização e da IA em contextos educacionais. Ainda, o trabalho feito por Preuss,

Barone e Henriques [11] investigou a aplicação de técnicas de IA em sistemas de ensino baseados em interfaces tangíveis, voltados especialmente para a educação inclusiva de crianças com deficiência intelectual ou autismo. O referente estudo destaca a forma de como a IA pode ser usada para personalizar trilhas de aprendizado para atender as necessidades individuais dos alunos, de modo a facilitar a supervisão da aprendizagem, além da gestão pedagógica. Com estas tecnologias, o sistema consegue avaliar o progresso dos alunos, adaptando o nível de desafio e fornecendo *feedback*, tornando o processo de aprendizagem inclusivo e personalizado.

A leitura e a análise da literatura indicam uma tendência crescente sobre o uso da IA como suporte ao ensino, o que destaca a importância do desenvolvimento de ferramentas que auxiliem não apenas no processo de aprendizado, mas que também sejam capazes de se adaptar e evoluir diante das necessidades educacionais atuais. No contexto das ferramentas apresentadas na literatura, a solução proposta neste trabalho se posiciona como uma alternativa voltada especificamente para o suporte educacional em cenários que envolvem dificuldades de aprendizado interdisciplinares.

Assim, este trabalho apresenta uma proposta que se insere nesse contexto, uma vez que a proposta pode ser aplicável a diversas áreas de estudo, ampliando sua acessibilidade e efetividade por meio do uso da IA. Diferentemente de outras abordagens, que muitas vezes focam em aplicações genéricas, o presente sistema combina três mecanismos (Verificação Rápida, Expansão e Correção de Erros) para oferecer uma experiência mais adaptativa e personalizada. O uso de técnicas como TF-IDF e similaridade de cosseno para otimizar o tempo de resposta, aliado a um modelo de linguagem treinado em português brasileiro (Sabiá 2.0), permite maior aderência ao contexto cultural e linguístico do público-alvo.

3 Arquitetura do Sistema

Esta seção descreve os elementos que compõem o sistema. A arquitetura do sistema pode ser vista na Figura 1. Detalhes de cada elemento são descritos a seguir:

- **Cliente**: O cliente representa a interface em que o usuário pode fazer perguntas relacionadas a diversas áreas do conhecimento, como Computação, Ciências Exatas e Humanas (conforme Figura 2a). O usuário interage com o sistema de forma intuitiva, submetendo perguntas conceituais, contextuais, comparativas e explicação detalhada. Algumas limitações relacionadas aos vieses de perguntas que exigem opinião subjetiva em que o modelo não possui a capacidade de responder, sendo então descartadas pelo sistema. Por fim, existe uma interface em que o usuário poderá ver seu histórico de perguntas (Figura 2b) caso seja necessário, evitando realizar uma requisição ao servidor.
- **Servidor e BD**: Ao receber uma pergunta, o sistema aciona o mecanismo de verificação rápida (seção 3.1.2) para verificar se já existe uma pergunta correspondente armazenada no banco de questões. O processo de verificação é realizado utilizando um método de similaridade de cosseno, que compara a nova pergunta com todas as perguntas previamente carregadas do banco de questões no início da aplicação. Por exemplo, se a pergunta for "Qual a importância da Inteligência Artificial no mundo moderno?", o sistema primeiro vetorizaria essa pergunta junto com as perguntas carregadas, e calcularia a similaridade entre elas. Se for encontrada uma pergunta com um índice de similaridade maior que o limiar predefinido (neste trabalho foi adotado 80%), o sistema retornará a resposta já existente de forma quase imediata. A Figura 1 mostra o fluxo descrito anteriormente, desde a pergunta do usuário até a apresentação da resposta ou chamada da API.

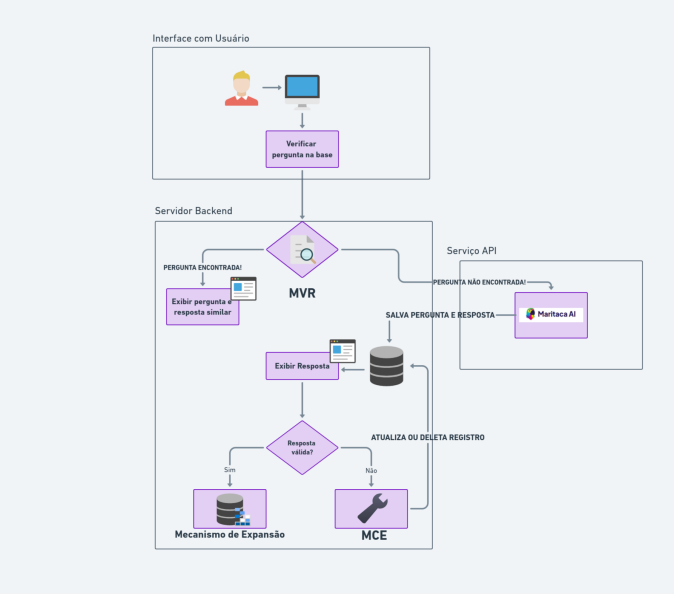
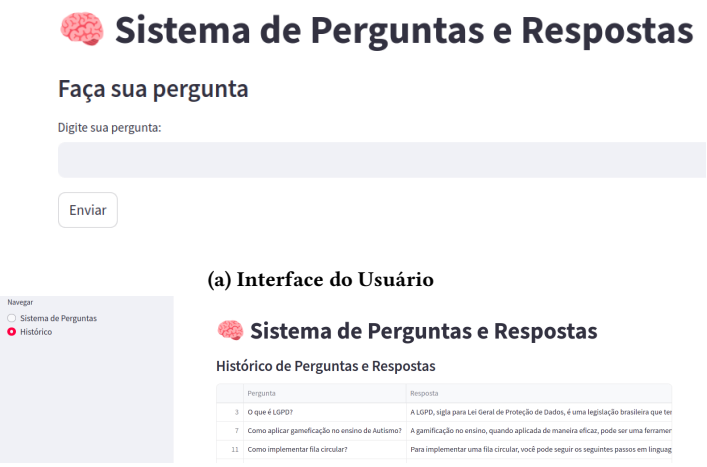


Figura 1: Arquitetura do Sistema



(a) Interface do Usuário

(b) Histórico de Perguntas

Figura 2: Interfaces do Sistema de Perguntas e Respostas

- **API Maritaca**: Caso o sistema não encontre uma pergunta similar como descrito no item anterior, ele acionará o processo que faz uma chamada usando a API Maritaca. Essa API é utilizada para gerar uma resposta para a pergunta realizada, adotando o modelo de linguagem Sabiá 2.0 para interpretar e responder a questão de maneira correta e contextualizada. Além disso, após gerar a resposta, o sistema adiciona automaticamente essa pergunta e a resposta correspondente ao banco de questões. Ainda, caso a resposta seja validada pelo usuário, o sistema adota a estratégia de aumentar a variabilidade do banco de questões, acionando o mecanismo de expansão, em que a partir da pergunta original são geradas diferentes formulações com suas respectivas respostas de modo a proporcionar maior robustez ao sistema. A Figura 1 mostra o processo do funcionamento do mecanismo de expansão.

3.1 Componentes Funcionais do Sistema

Na arquitetura do sistema existem três componentes funcionais chamados de Mecanismo de Expansão, Mecanismo de Verificação Rápida (MVR) e Mecanismo de Correção de Erros (MCE). Esses elementos fazem parte da arquitetura modular do sistema e são implementados como módulos mutuamente dependentes que trabalham em conjunto para melhorar a qualidade das respostas.

3.1.1 Mecanismo de Expansão. O mecanismo de expansão proposto neste estudo visa aumentar a diversidade do banco de questões, bem como aprimorar a precisão e a robustez do sistema. Utilizando API da Maritaca [10] com modelo de linguagem Sabiá 2.0, o servidor realiza chamadas para gerar diferentes versões de uma mesma pergunta e suas respectivas respostas. Este processo permite melhorar o banco de questões com novas perguntas baseadas em variações semânticas das originais, otimizando o sistema de perguntas e respostas. A sequência de ações do mecanismo de expansão pode ser vista na Figura 3.

Inicialmente, o banco de dados contém um conjunto de n perguntas, e cada uma delas pode ser expandida em k versões utilizando a API, resultando em um banco de perguntas com $n \times k$ entradas. Cada nova versão da pergunta é cuidadosamente criada pela API, que possui a capacidade de compreender nuances e variações na formulação. Isso maximiza as chances de o sistema oferecer respostas mais completas e precisas, mesmo para perguntas formuladas de maneiras bem diferentes. A configuração da requisição para a geração das respostas usada na API utiliza parâmetros que permitem controle sobre o comportamento do modelo de linguagem ao processar a entrada e gerar respostas. A Tabela 1 mostra detalhes da configuração que permite gerar respostas que são equilibradas em termos de coerência e criatividade, otimizando a interação no contexto de perguntas e respostas. Os parâmetros de configuração do modelo foram definidos com base em ajustes empíricos, realizados por meio de experimentos iniciais com o sistema. Esses experimentos buscaram equilibrar relevância e eficiência no contexto educacional. Embora esses ajustes tenham se mostrado adequados para o contexto específico deste estudo, os autores reconhecem que a definição ideal desses parâmetros é uma questão complexa que pode variar conforme o domínio ou o público-alvo.

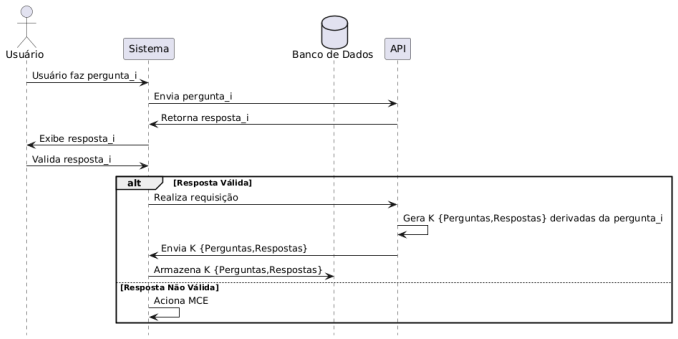


Figura 3: Mecanismo de Expansão

Tabela 1: Configuração dos Parâmetros de Geração de Respostas

Parâmetro	Valor	Descrição
do_sample	True	Ativa a amostragem aleatória, aumentando a variabilidade e a criatividade das respostas.
max_tokens	800	Define o limite máximo de *tokens* na resposta, equilibrando tempo de processamento e completude.
temperature	0.7	Controla a criatividade, em que 0.7 equilibra precisão e diversidade.
top_p	0.95	Limita as escolhas a um conjunto com probabilidade acumulada de 95%, melhorando a qualidade da resposta.

Considerando que o sistema recebeu a pergunta "O que é a LGPD?", o mecanismo de expansão pode gerar variações como "Qual a importância da LGPD para a privacidade de dados?" ou "Quais são os principais aspectos da LGPD no Brasil?", mantendo a coerência semântica, mas oferecendo uma gama maior de respostas possíveis. O objetivo desse mecanismo é melhorar a eficiência do sistema de Q&A ao lidar com perguntas diversas sem comprometer a qualidade das respostas.

3.1.2 Mecanismo de Verificação Rápida. O MVR implementado tem como objetivo otimizar o tempo de resposta do sistema. Ele atua de forma eficiente ao verificar se a pergunta feita pelo usuário já existe no banco de questões ou se há uma pergunta similar. O respectivo processo pode ser visto na Figura 4. A seguir segue o passo a passo do processo:

(1) **Entrada da Pergunta**: O usuário insere uma pergunta no sistema, que é imediatamente comparada com as perguntas previamente carregadas da base de questões.
(2) **Verificação de Similaridade**: O mecanismo utiliza a técnica TF-IDF (*Term Frequency-Inverse Document Frequency*) como sistema de pesos para medir a relevância das palavras de uma pergunta em relação as palavras armazenadas no banco de dados. Essa abordagem transforma o texto em representações numéricas, combinando duas etapas: (1) a vetorização dos textos, que converte as palavras em um espaço vetorial, e (2)

a atribuição de pesos as palavras com base nos valores de TF-IDF, indicando sua importância no contexto analisado. Após a vetorização, o sistema utiliza a métrica de similaridade de cosseno para comparar a pergunta do usuário com as perguntas existentes na base de dados.

(3) **Exibição do Resultado**: Caso uma pergunta com alta similaridade seja encontrada (acima de um limiar definido, por exemplo, 0.8), a resposta correspondente é exibida ao usuário conforme mostrado na Figura 5. Caso contrário, o sistema faz uma consulta usando a API Maritaca.

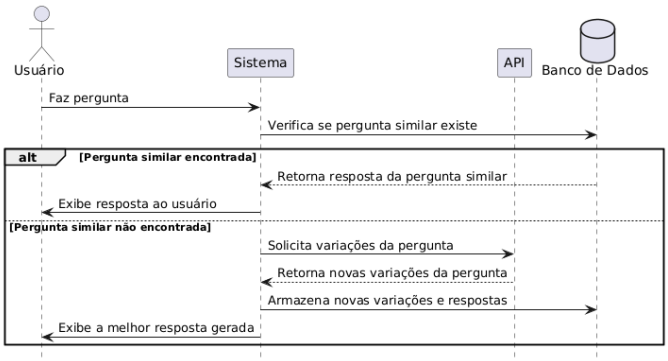


Figura 4: Mecanismo de Verificação Rápida (MVR)

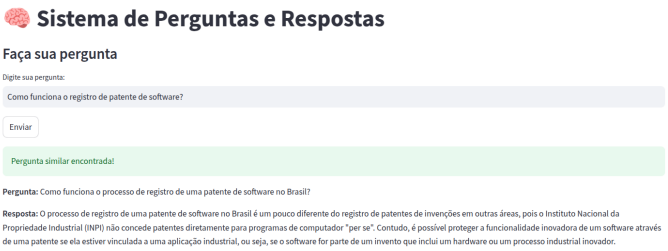


Figura 5: Exemplo de Interface do Usuário para o MVR

3.1.3 Mecanismo de Correção de Erros. O MCE é um dos componentes importantes na validação e melhoria da confiabilidade dos sistemas de perguntas e respostas assistidos por IA. Em um contexto educacional, é importante garantir que as respostas fornecidas sejam adequadas para permitir que os usuários confiem no sistema e também incentivar a melhoria contínua da qualidade das interações. No presente trabalho, a implementação do MCE buscou corrigir as respostas inconsistentes ou erradas. O processo de correção começa com o *feedback* do usuário, que pode informar se as respostas estão incorretas, desatualizadas ou insuficientes enquanto interage com o sistema. Uma vez relatado o erro, o sistema passa por uma série de etapas para apresentar uma melhor resposta ao mesmo tempo em que melhora seu banco de questões.

Uma vez que o erro é reportado, o sistema faz uma requisição para gerar uma resposta revisada. Esta nova resposta é então comparada com a anterior e, se validada pelo usuário, substitui a resposta incorreta e registra no banco de questões (conforme mostra a Figura 6). Caso contrário, o sistema descarta os dados. Esse processo não só garante a correção de erros em tempo real, mas também alimenta o modelo com *feedback* que pode ser utilizado para melhorias nas futuras interações.

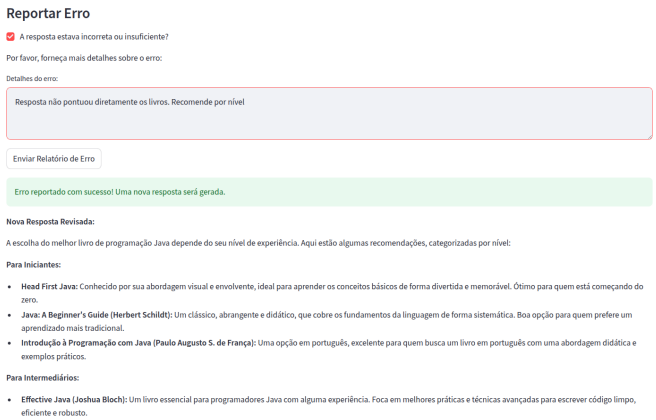


Figura 6: Exemplo de Interface do Usuário para o MCE

A Figura 7 ilustra cada etapa do processo do MCE em ação, desde o reporte do erro até a geração da nova resposta. A implementação do MCE neste estudo segue uma abordagem iterativa de aprendizado, na qual o *feedback* dos usuários não só corrige falhas, mas também serve como um mecanismo dinâmico de aprendizado para o sistema. Além disso, o armazenamento das respostas incorretas em uma base de dados separada permite que esses dados sejam revisados manualmente por um especialista ou utilizados para análise posterior, contribuindo para a melhoria contínua do sistema. Isso reflete a importância de ter um mecanismo adaptável, capaz de refinar seu desempenho com base nas experiências dos próprios usuários, aprimorando a eficiência e a confiabilidade do assistente de ensino.

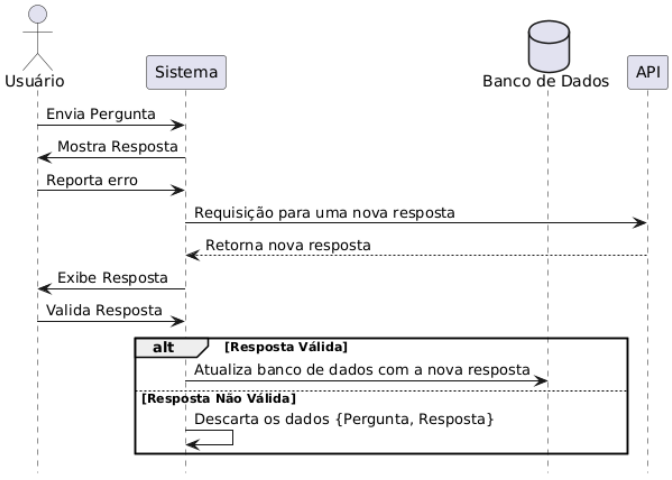


Figura 7: Mecanismo de Correção de Erros (MCE)

4 Metodologia

4.1 Abordagem de Pesquisa

O presente estudo adota a abordagem de *Design Science Research* (DSR) para integrar métodos quantitativos e qualitativos na avaliação de um sistema de Q&A assistido por IA no contexto educacional [14, 12]. A abordagem quantitativa visa quantificar o desempenho do sistema, particularmente por meio do tempo de resposta, enquanto a abordagem qualitativa busca interpretar as percepções dos alunos sobre a eficiência e a qualidade das respostas fornecidas. A posição epistemológica do estudo é uma combinação de positivismo, utilizado para obter medições objetivas e replicáveis da eficiência do sistema de IA, e de uma perspectiva interpretativa, para melhor compreender as reações e percepções dos alunos.

O DSR permite que o estudo combine a criação e a avaliação de um assistente de ensino a partir da coleta de dados, que são usados para avaliar sua eficácia. O método principal é o experimento, que possibilita uma observação direta do comportamento do sistema em resposta as interações dos alunos, permitindo uma validação prática das funcionalidades propostas e o refinamento do sistema com base nos resultados observados.

4.2 Técnicas de Coleta e Análise de Dados

Foram utilizadas duas principais técnicas de coleta de dados:

- **Registro de interações do usuário**: As perguntas submetidas e as respostas geradas foram armazenadas em um banco de dados para análise posterior.
- **Avaliação dos estudantes**: Os estudantes avaliaram as respostas geradas quanto ao tempo e a qualidade em uma escala de 1 a 5. Esses dados foram coletados por meio de questionários aplicados após o uso do sistema.

Para analisar os dados, utilizou-se Análise Estatística a fim de avaliar o tempo e a qualidade das respostas, permitindo quantificar a eficiência do sistema. Além disso, foram utilizados gráficos para melhor interpretação dos dados.

4.3 Validação e Teste da Pesquisa

Para a execução do sistema proposto, os testes foram conduzidos em um notebook com Sistema Operacional Linux, versão Ubuntu 22.04, processador Intel® Core™ i3-1115G4 de 11ª geração, e 4 GB de memória RAM. Os testes foram feitos para refletir um cenário de uso cotidiano, com recursos de hardware modestos, o que permitiu avaliar o desempenho do sistema em condições que são comumente encontradas em contextos educacionais no Brasil. Desta forma, a configuração do ambiente foi selecionada para garantir que os resultados fossem aplicáveis a ambientes de ensino reais. A validação incluiu:

- **Testes de desempenho**: Medição dos tempos de resposta do sistema usando MVR e o modelo Sabiá 2.0.
- ***Feedback* dos usuários**: Os estudantes avaliaram tanto a velocidade quanto a qualidade das respostas. As respostas foram agrupadas e analisadas para verificar se o sistema conseguia atender satisfatoriamente as necessidades dos alunos, considerando tanto o tempo de resposta quanto a relevância do conteúdo.

5 Resultados e Discussões

Os dados coletados sobre o uso do sistema de perguntas e respostas foram analisados com base em duas variáveis: qualidade da resposta e tempo de resposta (MVR e o modelo Sabiá 2.0). As avaliações deste trabalho foram separadas em duas categorias, pelo Mecanismo de Resposta (seção 5.1) e Usuários (seção 5.2). A seguir serão apresentados os resultados em detalhes.

5.1 Avaliação dos Mecanismos de Resposta

Para avaliação do desempenho das respostas do sistema foi utilizado o tempo de resposta da modelo Sabiá 2.0 e o MVR durante as perguntas dos usuários. Desta forma, foram selecionadas 25 perguntas aleatórias feitas pelos usuários, aplicadas duas vezes para medir o tempo de resposta de cada técnica utilizada pelo sistema. Os resultados estão apresentados nas Figuras 8 e 9.

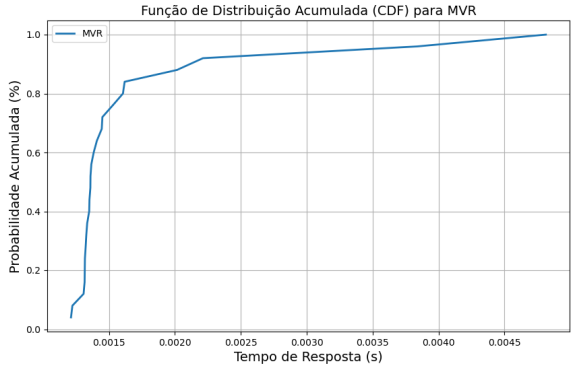


Figura 8: Distribuição do tempo de resposta do MVR

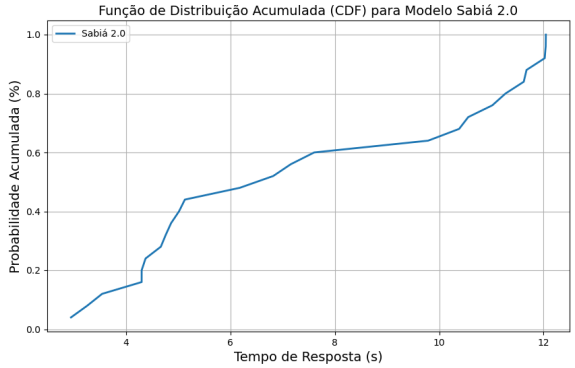


Figura 9: Distribuição do tempo de resposta do Sabiá 2

A Figura 8 apresenta a distribuição cumulativa dos tempos de resposta usando MVR. Foram observados os seguintes pontos:

- Conforme ilustra o gráfico, a probabilidade acumulada cresce rapidamente, indicando que uma grande porcentagem dos tempos de resposta (85%) são baixos, dentro do intervalo $[0.0015, 0.0020]$. Isso mostra que o MVR retorna de forma rápida a consulta na base de questões.

- A probabilidade acumulada se aproxima de 100% a medida que o tempo fica próximo de 0,0045 segundos, mostrando que maioria dos tempos de resposta estão abaixo desse valor.
- Esses resultados indicam que o tempo de resposta do MVR apresenta uma distribuição estável e restrita, mostrando que a distribuição dos tempos de repostas está concentrada em valores baixos.

A Figura 9 mostra a distribuição cumulativa dos tempos de resposta do Sabiá 2.0 em segundos. Foram observados os seguintes pontos:

- Diferente do gráfico da Figura 8, em que no início a acumulação foi rápida, no gráfico da Figura 9 a probabilidade acumulada cresce de forma mais lenta, indicando uma maior variação nos tempos de resposta iniciais. Isso mostra que o modelo Sabiá 2.0 possui uma latência mais alta em seus tempos de resposta.
- A curva atinge 100% de probabilidade acumulada em torno dos 12 segundos, mostrando que todos os tempos de resposta estão abaixo deste valor. O impacto desse resultado implica que o modelo responde de maneira inconsistente, com alguns tempos de resposta bastante elevados.
- A curva mostra uma subida gradual e um comportamento disperso, o que indica que o modelo possui vários picos de tempo de resposta. Este comportamento pode estar relacionado as variações no processamento das requisições que estão relacionadas as complexidade das perguntas ou da carga do servidor utilizado.

5.2 Avaliação dos Usuários

As avaliações dos usuários foram realizadas por um conjunto de alunos ($n = 20$) que interagiram com o sistema, fornecendo notas de 1 (pior avaliação) a 5 (melhor avaliação) para as variáveis tempo e qualidade, como ilustrado na Figura 10.

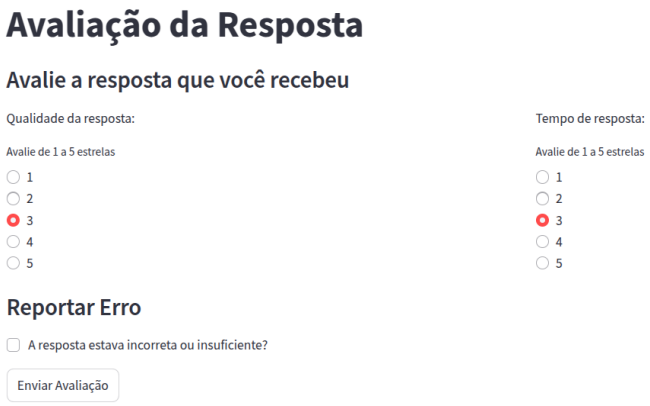


Figura 10: Interface de Avaliação das Respostas

Conforme a Figura 11, que apresenta a distribuição das avaliações da variável **qualidade**, foi observado que a maior parte das respostas recebeu notas entre 4 e 5, indicando um alto grau de satisfação por parte dos alunos. Assim, constatou-se que 60% das respostas foram avaliadas com a nota 5, mostrando que o sistema foi eficaz em gerar respostas coerentes e relevantes para as perguntas submetidas.

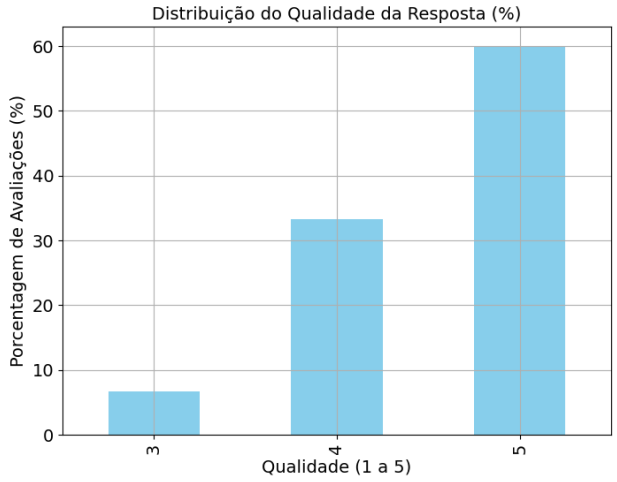


Figura 11: Distribuição das Avaliações da Qualidade das Respostas

Para efeitos estatísticos a média da qualidade das respostas foi de 4.67, com um desvio padrão de 0.72, indicando consistência na percepção de qualidade das respostas pelos alunos, refletindo precisão e coerência nas respostas fornecidas.

Conforme a Figura 12 que apresenta a distribuição das avaliações da variável **tempo**, foi observado que a maior parte das respostas recebeu notas entre 2 e 3, indicando um baixo grau de satisfação por parte dos alunos com relação ao tempo de resposta. Embora o sistema tenha apresentado avaliação positiva com relação a qualidade, houve uma certa insatisfação ao tempo de algumas resposta. Esse impacto negativo foi motivado pelas respostas geradas pelo modelo Sabiá 2.0 que demandam processamento adicional na geração de conteúdo.

Para efeitos estatísticos a média do tempo das respostas foi de 3.27, com um desvio padrão de 1.25, mostrando que o tempo de resposta é aceitável em muitos casos. Porém com cenários diferentes (usando outros modelos como Sabiá 3.0 ou ChatGPT 4.0) existe possibilidade de otimizar o tempo de resposta.

Para avaliar a satisfação dos usuários em relação ao tempo de resposta do sistema, foi realizado um agrupamento das avaliações de qualidade (**satisfação**) com base nas diferentes categorias de tempo de resposta, em que 5 representa o melhor tempo (resposta mais rápida) e 1 representa o pior tempo (resposta mais lenta). A Figura 13 apresenta a satisfação do usuário em relação de tempo de resposta.

Conforme mostrado na Figura 13, as respostas mais rápidas (avaliadas com 5) receberam valores elevados de satisfação (> 4), indicando uma correlação positiva entre a velocidade da resposta e a satisfação do usuário. A medida que o tempo de resposta reduz (4 para 2), observa-se uma diminuição gradual na satisfação. Esse comportamento mostra que os usuários tendem a considerar a qualidade da resposta mais satisfatória quando o sistema responde

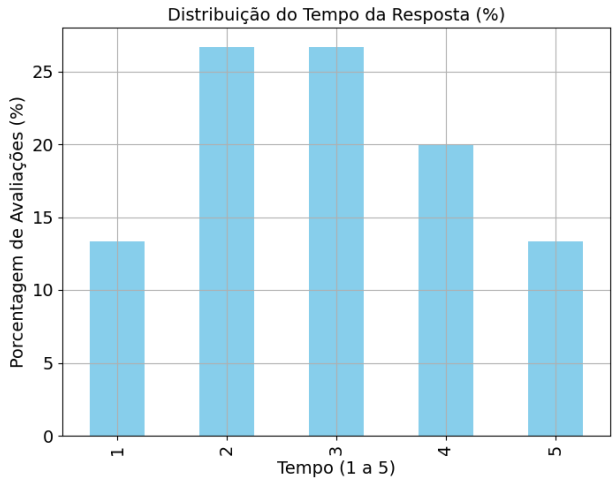


Figura 12: Distribuição das Avaliações do Tempo de Resposta

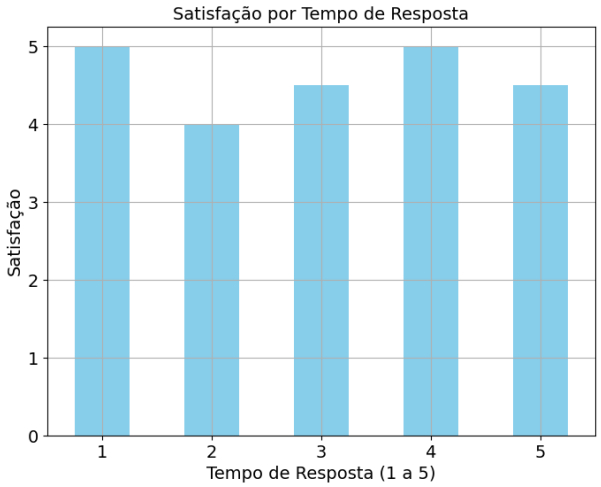


Figura 13: Distribuição da Satisfação pelo Tempo de Resposta

rapidamente. Esses resultados apresentados reforçam a importância de priorizar a rapidez do sistema, especialmente em consultas frequentes ou de alta demanda.

6 Conclusão

O presente trabalho apresentou o desenvolvimento e a avaliação de um assistente de ensino baseado em inteligência artificial, utilizando um modelo de linguagem brasileiro Sabiá 2.0. Durante os experimentos, o sistema foi avaliado com base na percepção dos usuários sobre a qualidade das respostas fornecidas e o tempo de resposta. Os resultados, como na Figura 13, indicaram alto grau de satisfação em termos de qualidade percebida, embora desafios relacionados ao tempo de resposta ainda estejam presentes, especialmente em consultas que exigem processamento mais complexo.

A implementação do Mecanismo de Verificação Rápida (MVR) e o Mecanismo de Expansão se destacaram ao otimizar o tempo de resposta e ampliar a abrangência do banco de questões. O MVR possibilitou respostas rápidas para questões similares, enquanto o Mecanismo de Expansão enriqueceu o sistema ao gerar variações semânticas das perguntas, permitindo maior flexibilidade no reconhecimento das dúvidas dos usuários. Além disso, o Mecanismo de Correção de Erros aprimorou a precisão das respostas ao longo do tempo, aumentando a confiabilidade do sistema.

Os resultados indicaram que, embora o sistema tenha obtido altos índices de satisfação em termos de qualidade das respostas, alguns desafios relacionados ao tempo de resposta permaneceram, especialmente em consultas processadas pelo modelo Sabiá 2.0. Esses desafios reforçam a importância de explorar soluções que mantenham a qualidade das respostas, mas com tempo de processamento reduzido.

Como direcionamentos para trabalhos futuros, recomenda-se o aprimoramento dos mecanismos de correção e verificação rápida, utilizando rede neural, bem como a implementação de versões mais avançadas do modelo (como chatGPT 4.0 ou Sabiá 3.0), visando otimizar o desempenho e elevar ainda mais a satisfação dos usuários. Além disso, a realização de testes A/B poderia avaliar sistematicamente diferentes combinações de parâmetros do modelo, de modo a identificar aquelas que otimizem o desempenho do sistema em termos de qualidade e tempo de resposta. Esse tipo de análise permitiria observar o impacto de cada configuração do modelo em termos globais, promovendo uma maior generalização e aplicabilidade do sistema em diversos contextos. Por fim, o respectivo trabalho confirma o potencial da inteligência artificial na área educacional, oferecendo uma solução que une acessibilidade e personalização no suporte ao aprendizado, contribuindo para o avanço do ensino assistido por IA.

Pesquisadores e profissionais que desejem explorar, reproduzir ou colaborar no aprimoramento deste projeto podem acessar o código-fonte completo disponibilizado no repositório GitHub conforme link: https://github.com/eltonsarmanho/AssistenteEnsinoUFPA_Cameta.

7 Agradecimentos

Os autores agradecem a referida empresa **Maritaca AI** [10] pelo fornecimento gratuito de acesso ao seu *Large Language Model* (LLM), o que possibilitou a realização deste estudo. O suporte tecnológico oferecido foi essencial para a implementação e validação das soluções desenvolvidas, contribuindo significativamente para os avanços apresentados neste trabalho.

Referências

[1] Imran Ahmed, Gwanggil Jeon e Francesco Piccialli. 2022. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18, 8, 5031–5042. DOI: 10.1109/TII.2022.3146552.

[2] Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira e Ramon Pires. 2024. Sabiá-2: a new generation of portuguese large language models. (2024). https://arxiv.org/abs/2403.09887 arXiv: 2403.09887 [cs.CL].

[3] Yiqian Chen, Hanxi Deng, Chi-Hua Chen e Chan-Liang Chung. 2023. Efficient artificial intelligence-teaching assistant based on chatgpt. Em *2023 International*

Conference on Smart Systems for applications in Electrical Sciences (ICSSES), 1–5. DOI: 10.1109/ICSSES58299.2023.10200077.

[4] Yingjun Gong. 2021. Application of virtual reality teaching method and artificial intelligence technology in digital media art creation. *Ecol. Informatics*, 63, 101304. https://api.semanticscholar.org/CorpusID:234817655.

[5] Yann Hicke, Anmol Agarwal, Qianou Ma e Paul Denny. 2023. Ai-ta: towards an intelligent question-answer teaching assistant using open-source llms. (2023). https://arxiv.org/abs/2311.02775 arXiv: 2311.02775 [cs.LG].

[6] Stefan Huber, Kristian Kiili, Steve Nebel, Richard Ryan, Michael Sailer e Manuel Ninaus. 2024. Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, 36, (fev. de 2024). DOI: 10.1007/s10648-024-09868-z.

[7] Jaeho Jeon e Seongyong Lee. 2023. Large language models in education: a focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, (mai. de 2023), 0000. DOI: 10.1007/s10639-023-11834-1.

[8] Jinhee Kim, Hyunkyung Lee e Young Cho. 2022. Learning design to support student-ai collaboration: perspectives of leading teachers for ai in education. *Education and Information Technologies*, 27, (jan. de 2022), 1–36. DOI: 10.1007/s10639-021-10831-6.

[9] Treice Moreira, Cláudia Silva, Cláudio Passos, Isabel Fernandes e Ronaldo Goldschmidt. 2023. Tutor inteligente em jogo educacional digital para capacitação na identificação de fake news em português: experimentos preliminares. Em *Anais Estendidos do XXXIV Simpósio Brasileiro de Informática na Educação*. SBC, Passo Fundo/RS, 14–20. DOI: 10.5753/sbie_estendido.2023.235269.

[10] Ramon Pires, Hugo Abonizio, Thales Sales Almeida e Rodrigo Nogueira. 2023. [inline-graphic not available: see fulltext] sabiá: portuguese large language models. Em *Intelligent Systems: 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25–29, 2023, Proceedings, Part III*. Springer-Verlag, Belo Horizonte, Brazil, 226–240. ISBN: 978-3-031-45391-5. DOI: 10.1007/978-3-031-45392-2_15.

[11] Evandro Preuss, Dante Barone e Renato Henriques. 2020. Uso de técnicas de inteligência artificial num sistema de mesa tangível. Em *Anais do XXVI Workshop de Informática na Escola*. SBC, Evento Online, 439–448. DOI: 10.5753/cbie.wie.2020.439.

[12] James P. Smith. 2015. A case study on design science research as a methodology for developing tools to support lean construction efforts. English. Em *23rd Annual Conference of the International Group for Lean Construction*. Olli Seppänen, Vicente A. González e Paz Arroyo, editores. Perth, Australia, 517–526. http://www.iglc.net/papers/details/1161.

[13] Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu e Philip Yu. 2024. Large language models for education: a survey. (Mai. de 2024). DOI: 10.48550/arXiv.2405.13001.

[14] Fernando Zaidan, Marcello Bax e Fernando Silva Parreiras. 2016. Design science research: aplicação em um projeto de pesquisa e desenvolvimento. Em *13th International Conference on Information Systems and Technology Management - CONTECSI*. (Jun. de 2016), 3757–3774. DOI: 10.5748/9788599693124-13CONTECSI/PS-4163.