Geração de Questões com LLMs Leves: Um Estudo Inicial sobre a Percepção de Educadores

Mateus Monteiro Santos [1], Aristoteles Peixoto Barros[1], Ermesson Santos [1], Jardilene Gomes da Silva [1], Seiji Isotani [3], Ig Ibert Bittencourt [1], Valmir Macario [4], Luiz Rodrigues[2], Diego Dermeval [1]

[1] Núcleo de Excelência em Tecnologias Sociais (NEES) – Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brasil

[2]Universidade Tecnológica Federal do Paraná (UTFPR) – Campus Apucarana
Apucarana – PR – Brasil

[3]Universidade da Pensilvânia (UPenn)
Filadélfia – PA – EUA

[4]Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

mateus.monteiro@nees.ufal.br, aristoteles.barros@nees.ufal.br

ermesson.santos@nees.ufal.br, diego.matos@nees.ufal.br

luiz.rodrigues@utfpr.edu.br

Abstract. *LLMs can assist in the automatic generation of questions but require internet access or robust infrastructure, which poses a barrier in low-resource settings. Lightweight LLMs, capable of running offline on regular smartphones, emerge as an alternative, although evidence regarding the quality of the generated content is still limited. This study investigates the use of such models to generate math questions. A total of 59 questions were produced and evaluated by a pedagogue, yielding mixed results. Most were considered usable, especially those involving addition and subtraction, but showed conceptual, grammatical, and semantic limitations. The study highlights the potential of lightweight LLMs and the need for human review to ensure pedagogical quality.*

Resumo. *LLMs podem auxiliar na geração automática de questões, mas exigem internet ou infraestrutura robusta, um obstáculo em contextos com poucos recursos. LLMs leves (lightweight), que operam offline em smartphones comuns, surgem como alternativa, embora ainda faltem evidências sobre a qualidade dos conteúdos gerados. Este trabalho investiga o uso desses modelos para gerar questões de matemática. Foram produzidas 59 questões e avaliadas por uma pedagoga, com resultados mistos. A maioria mostrou-se utilizável, sobretudo em operações de soma e subtração, mas com limitações conceituais, gramaticais e semânticas. O estudo destaca o potencial dos LLMs leves e a necessidade de revisão humana para garantir qualidade pedagógica.*

DOI: 10.5753/sbie.2025.12574

1. Introdução

O ensino de matemática ocupa papel central na formação de competências essenciais para a vida em sociedade, como o pensamento crítico, a capacidade de resolver problemas e o raciocínio lógico. Entretanto, a promoção de uma aprendizagem significativa nessa área permanece um desafio na educação básica brasileira, como evidenciam os resultados em avaliações nacionais e internacionais de larga escala, como SAEB e PISA [OECD 2023, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) 2022]. Esses resultados revelam dificuldades de desempenho e ressaltam a necessidade de soluções para apoiar estudantes e professores.

Tecnologias educacionais baseadas em inteligência artificial, como os Sistemas Tutores Inteligentes (STIs), têm se destacado por proporcionar instrução personalizada e feedback adaptativo, ajustando-se ao desempenho de cada estudante [Vanlehn 2006, Xu et al. 2019]. No entanto, a ampla adoção dessas tecnologias ainda enfrenta barreiras, sobretudo em escolas públicas e regiões remotas, devido à necessidade de infraestrutura tecnológica (e.g., computadores para os estudantes interagirem com os STIs) e conectividade constante (e.g., muitas soluções são baseadas em internet) [Gasevic et al. 2018, Soofi and Ahmed 2019].

Nos últimos anos, abordagens baseadas no paradigma de Inteligência Artificial na Educação (AIED) desconectada[1] [Isotani et al. 2023] vêm ganhando espaço por possibilitarem a utilização de recursos de IA em ambientes com poucos recursos tecnológicos, aproveitando dispositivos acessíveis como smartphones e contando com a mediação ativa dos professores [Freitas et al. 2022, Portela et al. 2023, Veloso et al. 2023]. Propostas como o MathAIde, um aplicativo móvel baseado nos princípios de STIs, demonstram avanços nesse sentido ao permitir a avaliação e a geração de listas de atividades matemáticas alinhadas à Base Nacional Comum Curricular (BNCC), mesmo em contextos de acesso restrito à tecnologia [Rodrigues et al. 2024b]. Assim como outras soluções baseadas em AIED desconectada, professores podem acessar o MathAIde com um dispositivo amplamente disponível atualmente (i.e., smartphones) para explorar os benefícios de STIs mesmo na falta de computadores, uma vez que o aplicativo permite gerar listas de atividades personalizadas, imprimi-las, coletar a resoluções feitas por estudantes com papel e lápis, e avaliar e fornecer feedback personalizado para tais atividades após coletá-las por meio da câmera do celular [Rodrigues et al. 2024a].

Apesar do potencial, a escalabilidade de soluções como o MathAIde permanece desafiadora. Em particular, o MathAIde conta com uma base pré-definida de questões, gerando limitações em termos da diversidade e demandas de atualização do banco de questões [Rodrigues et al. 2024a]. Em um contexto mais amplo, para além do paradigma de AIED desconectada, um dos principais desafios de STIs é a manutenção do componente pedagógico, que muitas vezes armazena as atividades educacionais que serão realizadas por seus usuários. Além disso, mesmo em contextos de ensino não baseados em

[1]Note que o paradigma de AIED desconectada difere do conceito *desplugado*, muitas vezes usado no ensino de computação. Ao contrário do ensino de computação desplugado, por exemplo, o paradigma de AIED desconectado não prevê soluções completamente offline ou independentes de tecnologias. O AIED desconectado fomenta soluções alinhadas aos recursos disponíveis e público-alvo, tirando vantagem a infraestrutura disponível em vez de exigir a aquisição de novos recursos (e.g., compra de computadores) [Isotani et al. 2023].

tecnologia, dada a necessidade de variar e personalizar atividades para turmas grandes e heterogêneas, professores destacam o desafio de encontrar materiais relevantes para suas aulas [Rodrigues et al. 2023, Guerino et al. 2024]

Os avanços em Grandes Modelos de Linguagem (LLMs) abre novas possibilidades para enfrentar as limitações de disponibilidade e acesso a recursos educacionais como questões / atividades. Ao permitir a geração automática de questões (ou atividades), os LLMs abrem caminho para maximizar a sustentabilidade de soluções como o MathAIde com uma quantidade pseudo-infinita de opções. Similarmente, o uso direto desses modelos pode empoderar professores na preparação de suas aulas. Porém, o acesso a esses modelos é tradicionalmente feito via internet (e.g., interface web de ChatGPT ou DeepSeek), pois exigem altos volumes de processamento. Embora isso gere restrições de acesso a público com recursos tecnológicos limitados, estudos recentes têm explorado o uso de LLMs de maneira offline, executando-os diretamente em smartphones. Isso é possível graças a técnicas de distilação e quantização, que reduzem o tamanho total dos modelos e, consequentemente, os recursos necessários para utilizá-los. Dessa forma, os LLMs podem fornecer apoio pedagógico em contextos com recursos limitados, seja por meio de uso direto (e.g., professores interagem com o LLM) ou integrado em outras soluções (e.g., como o MathAIde).

Apesar do potencial de executar LLMs offline em dispositivos móveis, desafios práticos ainda limitam a ampla adoção dessas soluções em ambientes educacionais. Como identificado em [Monteiro Santos et al. 2025], alguns desses desafios incluem o consumo de memória e o tempo de resposta desses modelos. Além disso, os autores também destacaram a necessidade de estudos que vão além dos aspectos técnicos e exploram a qualidade do conteúdo gerado por LLMs otimizados para uso offline em smartphones. Nesse contexto, existem pesquisas analisando a geração de questões com modelos como o ChatGPT, majoritariamente na língua inglesa, reforçando o potencial de utilizar LLMs para equipar STIs e professores na preparação de aula [Sewunetie and Kovács 2023, Indran et al. 2024, Lee et al. 2024]. No entanto, para o melhor de nosso conhecimento, pesquisas anteriores não investigaram a qualidade dos conteúdos gerados em português brasileiro por LLMs otimizados para execução offline em smartphones. Logo, existe uma lacuna empírica sobre a capacidade desses modelos em gerar questões / atividades educacionais relevantes para o ensino de matemática no contexto brasileiro.

Em direção à preencher a lacuna de pesquisa citada acima, neste trabalho em andamento, investigamos o uso de LLMs leves para a geração automática de questões matemáticas no contexto da educação básica brasileira, com foco em ambientes offline e alinhamento à BNCC. O objetivo é analisar os desafios e potencialidades pedagógicas dessa abordagem, contribuindo para a discussão sobre caminhos para uma educação matemática mais equitativa e de qualidade em escolas com infraestrutura restrita.

2. Revisão da Literatura

O uso de STIs tem sido amplamente reconhecido como uma das abordagens mais eficazes para a personalização do ensino de matemática, combinando inteligência artificial e análise de dados para oferecer instrução e feedback individualizados [Vanlehn 2006]. Diversas pesquisas e meta-análises comprovam o impacto positivo dos STIs na aprendi-

zagem matemática em diferentes níveis de ensino [Xu et al. 2019, Hillmayr et al. 2020]. No entanto, grande parte desses avanços permanece restrita a ambientes com boa infraestrutura tecnológica, limitando seu alcance em contextos com poucos recursos e, consequentemente, ampliando desigualdades educacionais [Soofi and Ahmed 2019, Veloso et al. 2023].

Diante dessas limitações, o paradigma de AIED desconectado surge como resposta ao desafio de viabilizar tecnologias educacionais baseadas em inteligência artificial em regiões desconectadas ou com acesso limitado à internet [Isotani et al. 2023]. Nesse contexto, soluções que associam processamento local em dispositivos móveis, armazenamento offline e fluxos de trabalho centrados no professor têm buscado mitigar a dependência de conectividade contínua e ampliar o acesso à educação de qualidade [Portela et al. 2023, Freitas et al. 2022]. Em particular, a geração automática de questões matemáticas surge como elemento fundamental nessas abordagens, pois permite adaptar atividades às necessidades dos estudantes mesmo em contextos offline.

Tradicionalmente, a geração de questões nesses sistemas depende de bancos de itens construídos manualmente por especialistas, o que limita a escalabilidade e a diversidade das atividades [Rodrigues et al. 2024a, Rodrigues et al. 2024b]. Apesar dos avanços de soluções de AIED desconectada como o MathAIde, que combinam bancos de itens alinhados à BNCC, rastreamento de conhecimento leve e análise automatizada das respostas dos estudantes para apoiar o planejamento docente e fornecer feedback imediato, o esforço manual necessário para a construção e manutenção desses bancos limita a abrangência da solução [Rodrigues et al. 2024a, Rodrigues et al. 2024b, Veloso et al. 2023]. Assim, torna-se essencial investigar métodos automáticos que ampliem a escala da geração de questões sem comprometer a qualidade pedagógica.

Diante desse cenário, estudos sobre a adoção local de novas tecnologias para geração automática de questões matemáticas representam um passo fundamental para potencializar a personalização, ampliar a equidade e garantir a viabilidade técnica de soluções educacionais inteligentes em ambientes desconectados. Notavelmente, a geração automática de questões tem sido investigada em diferentes pesquisas. [Lee et al. 2024] utilizaram uma abordagem baseada em pesquisa e desenvolvimento para propor um sistema de geração de questões que usa o ChatGPT e engenharia de prompt[2], alcançando evidências empíricas sobre a validade das questões geradas. Por sua vez, [Indran et al. 2024] apresentam dicas práticas para o uso do ChatGPT na geração de questões médicas, concluindo que a ferramenta tem forte potencial para apoiar educadores. Similarmente, [Sewunetie and Kovács 2023] discutem o potencial do ChatGPT, destacando a importância de garantir a qualidade do conteúdo geral.

Ou seja, enquanto a literatura demonstra o potencial e a relevância dos LLMs para geração de questões, estudos anteriores são focados no ChatGPT e na língua inglesa, deixando uma lacuna de conhecimento sobre a capacidade de modelos ajustados para uso em contextos de baixo recurso do Brasil. Além disso, observa-se que grande parte dessas investigações sobre engenharia de prompt permanece em caráter empírico, sem uma sistematização detalhada. Estratégias como a atribuição explícita de papéis ao modelo

[2]Engenharia de prompt refere-se ao processo de projetar e estruturar comandos (prompts) de forma estratégica, a fim de orientar modelos de linguagem a produzir respostas mais relevantes e adequadas ao contexto.

como por exemplo, instruí-lo a atuar como "um professor especialista em matemática para crianças de 7 anos" são apontadas pela literatura como técnicas eficazes para melhorar a clareza e o contexto das respostas, mas ainda pouco exploradas nesse domínio. Logo, este trabalho em andamento apresenta um passo em direção a preencher tal lacuna ao investigar a capacidade de um LLM leve, executado em um smartphone online, de gerar questões matemáticas relevantes.

3. Metodologia

Este estudo explora o uso de um LLM leve para a geração automática de questões matemáticas em ambientes desconectados (offline), por meio da execução local desses modelos em dispositivos móveis com sistema Android. Em particular, este estudo em andamento concentra-se em avaliar o comportamento de um LLM leve na geração de questões matemáticas para o ensino fundamental, considerando a qualidade desses conteúdos com base em aspectos como acurácia, completude, clareza e intenção de uso didático. Embora não se proponha a oferecer uma solução definitiva, a pesquisa busca contribuir para a compreensão inicial sobre os desafios e potencialidades envolvidas na adoção local de LLMs em contextos educacionais com recursos restritos.

3.1. Arquitetura e Execução Local do LLM

O modelo utilizado neste estudo foi o `gemma-2-2b-it-IQ3_M_imat.gguf` [Team 2024], um LLM de pequeno porte, quantizado no formato `.gguf`, e compatível com execução offline por meio da biblioteca `llama.cpp` [Gerganov 2023]. O modelo foi integrado em um aplicativo Android customizado, permitindo a geração de respostas diretamente no dispositivo, sem necessidade de conexão à internet. O ambiente de execução consistiu em um smartphone Samsung Galaxy A34, com sistema Android 14, escolhido por representar uma configuração intermediária com ampla distribuição global [StatCounter 2024]. A viabilidade de executar LLMs nesse dispositivo já havia sido demonstrada em estudo anterior [Monteiro Santos et al. 2025], o que reforça sua adequação para testes em cenários educacionais com infraestrutura limitada. O aplicativo foi desenvolvido em React Native, com a engine de inferência em C++ embutida via NDK, garantindo portabilidade e desempenho satisfatório.

A escolha do `gemma-2-2b-it` deve-se ao fato de que, embora existam outros LLMs leves quantizados disponíveis (como versões reduzidas de Phi, Mistral ou Llama), este modelo se mostrou mais adequado para execução em dispositivos com recursos tecnológicos limitados [Monteiro Santos et al. 2025], conciliando leveza computacional com qualidade de geração de respostas em tarefas educacionais.

3.2. Processo de Geração de Questões

O fluxo operacional foi estruturado em três etapas principais. Primeiro, a definição de um *prompt* base por tipo de operação matemática. Por exemplo, para questões de soma, utilizou-se o enunciado "*Usando menos do que vinte palavras, escreva uma situação problema para avaliar o conhecimento de uma criança de sete anos na seguinte habilidade: Questão do aluno: Soma*". Segundo, a geração de múltiplas instâncias de questões a partir desses *prompts*, utilizando o modelo LLM embarcado no dispositivo. Terceiro, a tabulação das questões geradas para avaliação posterior. Ao, foram produzidas 59 questões, distribuídas entre os quatro tipos básicos de operação (soma, subtração,

multiplicação e divisão), com variações mínimas nos prompts, mantendo consistência temática e operacional [3].

Dado que este é um estudo em andamento, optamos por não gerar um conjunto extensivo de questões, realizando uma análise inicial focada nos quatro tipos de operações básicas. Como resultado, essa abordagem possibilitou alcançar nosso objetivos de investigar o potencial de um LLM leve de forma robusta, com base na avaliação de uma pessoa especialista (veja a Seção 3.3), o que seria desafiador e demorado com um conjunto extenso de questões. Por fim, vale notar que a estrutura dos *prompts* foi definida de forma empírica, com base em observações diretas das respostas geradas, mas fundamentada em princípios discutidos na literatura sobre geração de questões com LLMs, especialmente quanto à importância de instruções claras, delimitação da habilidade e adequação à faixa etária [Scaria et al. 2024, Li and Zhang 2024, Nikolovski et al. 2025].

3.3. Critérios de Avaliação Qualitativa

A avaliação das questões foi conduzida por uma pedagoga, graduada em Licenciatura em Matemática pela Universidade Federal de Alagoas (UFAL), com experiência docente na rede privada de ensino em Maceió, onde lecionou matemática para turmas dos anos iniciais e finais do Ensino Fundamental ao longo de dois anos. Para estruturar a análise, adotamos um formulário estruturado com base em princípios de pesquisa qualitativa em educação, conforme sugerido por [Creswell and Creswell 2017] e por estudos sobre tutores inteligentes e avaliação de interações educacionais [Graesser et al. 2004, Vanlehn 2006]. Embora essas obras não abordem diretamente a avaliação qualitativa de respostas geradas por modelos de linguagem, oferecem fundamentos metodológicos relevantes para a definição de categorias e critérios de análise. Com base nisso, foram adotadas quatro dimensões de avaliação:

- **Acurácia**: verifica se a questão apresenta lógica matemática adequada e ausência de erros conceituais;
- **Completude**: avalia se todos os dados necessários para a resolução estão presentes e bem contextualizados;
- **Clareza**: considera a legibilidade, estrutura linguística e adequação à faixa etária;
- **Intenção de uso**: reflete o grau de aceitabilidade da questão para uso real em sala de aula.

Essas categorias foram selecionadas com base em sua relevância para contextos práticos, onde a *acurácia* contribui para os objetivos de aprendizagem, a *completude* garante a autonomia do estudante na resolução da questão e a *clareza* facilita o engajamento e a compreensão [Xin 2023]. Além disso, a verificação da *intenção de uso* fornece uma medida objetiva sobre a viabilidade prática de adoção das questões geradas pelo procedimento em questão [Venkatesh and Davis 2000]. Cada critério foi avaliado numa escala Likert de 1 a 5, e um campo adicional de comentários foi disponibilizado para observações qualitativas.

3.4. Análise dos Dados

A análise dos dados foi conduzida a partir de estatísticas descritivas, considerando medidas como média, mediana e desvio padrão para cada critério de avaliação, segmentadas

[3] A tabela com as 59 questões geradas, bem como os respectivos *prompts* utilizados, está disponível nesse link.

por tipo de operação matemática. Além disso, os comentários qualitativos fornecidos pela pedagoga foram examinados manualmente, com o objetivo de identificar padrões recorrentes de problemas linguísticos, estruturais ou conceituais. Essa etapa buscou complementar as análises quantitativas com percepções pedagógicas mais elaboradas, alinhadas às recomendações de estudos sobre avaliação humana de respostas educacionais em sistemas de IA [Graesser et al. 2004].

4. Resultados

A Tabela 1 apresenta os resultados quantitativos da avaliação. Em termos de acurácia, completude e legibilidade, os resultados são semelhantes, com valores médios variando entre 3 e 4 para os quatro tipos de operação. Vale destacar que o desvio padrão (DP) varia entre 1 e 2 para as três médicas, sugerindo uma variação notável na acurácia, completude e legibilidade das questões geradas. Por outro lado, a mediana é igual ou maior do que 4 na maioria das situações, indicando que a maioria das questões foi avaliada de forma satisfatória nesses critérios.

Além disso, a Tabela 1 também apresenta os resultados de intenção de uso. Em termos gerais, o valor médio ficou entre próximo do ponto central da escala Likert de cinco pontos, variando entre 2.67 e 3.47 com DPs que vão de 1.23 a 1.88. Diferente dos demais critérios, a mediana indica resultados claramente mistos. Para soma e subtração, a mediana foi 4, enquanto a mesma foi 1 e 2 para multiplicação e divisão, respectivamente. Esses resultados sugerem que, em termos de adoção prático, a maioria das questões de soma e subtração seriam adotadas, enquanto a maioria das questões de multiplicação e divisão seriam descartadas.

Tabela 1. Avaliação das questões geradas automaticamente, agrupadas por operação, em termos de média (M), mediana (MD) e desvio padrão (DP).

	Acurácia			**Completude**			**Legibilidade**			**Intenção de Uso**		
	M	MD	DP	M	MD	DP	M	MD	DP	M	MD	DP
/	3.64	3	1.22	3.71	3	1.27	3.14	3	0.95	2.71	2	1.38
$*$	3.13	5	2.07	3.13	5	2.07	3.00	3	2.00	2.67	1	1.88
+	3.20	4	1.57	3.53	5	1.73	3.47	5	1.77	3.47	4	1.68
-	3.33	4	1.23	3.93	5	1.83	3.20	4	1.37	3.33	4	1.23

A Tabela 2 exemplifica questões geradas para cada um dos tipos de operações básicas. A tabela também apresenta comentários a respeito de cada questão, ilustrando pontos de melhoria para cada um delas. Em particular, os insumos qualitativos indicaram que 14 das 59 (24%) questões estavam incorretas, por razões como explorar operações incorretas (e.g., questões de soma que deveriam ser de divisão), ausência de dados essenciais (e.g., indicar que valores devem ser divididos igualmente) e subjetividade dos enunciados. Dessa forma, enquanto os resultados quantitativos indicarem a adequação e potencial de uso prático de muitas questões, principalmente aquelas de soma e subtração, os resultados qualitativos contribuem insumos sobre limitações que devem ser endereçadas nas questões problemáticas.

Tabela 2. Exemplos de questões geradas pelo LLM e comentários da avaliadora

	Questão Gerada	Comentário da Avaliadora
/	João tem 12 balas para dividir com 3 amigos. Quantas cada um vai receber?	Ausência da palavra "igualmente" pode gerar ambiguidade no enunciado.
-	Ana tinha algumas balas e agora tem 3. Quantas ela perdeu?	Falta de dados suficientes para resolução matemática.
+	Pedro tem 5 laranjas e ganha mais 4. Quantas ele tem agora?	Correta, clara e adequada para a faixa etária do ensino fundamental.
*	Uma caixa de chocolates tem 8 caixas. Quantos chocolates há no total?	Redundância estrutural e confusão conceitual no uso do termo "caixas".

5. Discussão

A avaliação de questões geradas automaticamente com um LLM leve, compatível contextos com recursos limitados e o paradigma de AIED desconectado, revela resultados mistos. Por um lado, os resultados quantitativos indicam que a maioria das questões, é adequada em termos de acurácia, completude e legibilidade, principalmente aquelas focadas em operações de soma, subtração e multiplicação. Além disso, os resultados também sugerem que a maioria das questões de soma e subtração tem potencial para uso prático. Por outro lado, os resultados também revelam limitações claras, tanto com base no alto desvio padrão das avaliações quantitativas, quanto os insumos qualitativos. Particularmente, os resultados indicam que muitos dos enunciados avaliados apresentaram falhas conceituais, ambiguidade linguística ou estrutura inadequada para o público infantil. Assim, os achados dessa pesquisa em andamento contribuem para a literatura e a prática de diferentes formas.

Do ponto de vista prático, os resultados sugerem que o uso desse tipo de solução pode apoiar educadores por meio de um processo híbrido, mas não automação completa. Isso significa que o uso de modelos de linguagem leve para a geração de questões pode ser viável como uma ferramenta de apoio, desde que haja a atuação ativa do professor no processo de curadoria e validação das questões geradas. A presença do educador no loop é essencial para garantir que os conteúdos sejam pedagogicamente adequados, livres de ambiguidades e alinhados ao nível de desenvolvimento dos alunos. Esses insumos corroboram pesquisas anteriores, tanto no contexto de geração de questões [Indran et al. 2024, Lee et al. 2024] quanto na concepção de soluções de AIED [Rodrigues et al. 2023, Tenório et al. 2022], expandindo a literatura com insumos sobre o potencial de LLMs ajustados para contextos com recursos tecnológicos limitados.

Do ponto de vista de pesquisa, os resultados informam investigações no uso de tecnologias avançadas de IA em contextos de baixo recurso. Os achados desta pesquisa destacam a importância de avaliar criticamente a qualidade dos conteúdos gerados por LLMs antes de sua adoção prática em ambientes educacionais. Esses corroboram pesquisas sobre a importância de garantir que sistemas de AIED sejam confiáveis e alinhados a necessidades educacionais [Xia et al. 2022] a fim de garantir que tenham relevância prática. Logo, nossos achados orientam futuras investigações voltadas ao aprimoramento de métodos automáticos de geração de conteúdo, além de contribuir evidências para o uso de métricas alinhadas à qualidade educacional e estratégias de envolvimento do professor no ciclo de uso da IA.

Nesse contexto, este artigo em andamento fornecer duas implicações principais. Primeiro, embora o modelo avaliado tenha utilidade em tarefas iniciais, como rascunho de questões ou sugestões automáticas, não é recomendável integrá-lo diretamente em aplicativos educacionais que façam geração automática de conteúdo sem revisão humana, especialmente em contextos voltados para o público infantil, onde a clareza conceitual e a adequação linguística são fundamentais. Segundo, os resultados indicam a necessidade de pesquisas voltadas a entender como melhorar a qualidade dos conteúdos gerados por esses modelos, seja por meio de ajustes nos próprios algoritmos, uso de técnicas de pós-processamento, ou pela incorporação de mecanismos de avaliação automática mais sensíveis a critérios pedagógicos e linguísticos. Assim, enquanto a literatura apresenta investigações o desenho e avaliação de soluções alinhadas ao paradigma de AIED desconectada [Freitas et al. 2022, Portela et al. 2023, Veloso et al. 2023] ou a geração de questão em língua inglesa com LLMs que exigem acesso à internet ou recursos tecnológicos robustos [Lee et al. 2024, Sewunetie and Kovács 2023, Indran et al. 2024], este artigo se difere ao contribuir evidências sobre o potencial de um LLM leve para geração de questões em português brasileiro.

6. Limitações e Trabalhos Futuros

Este estudo apresenta quatro limitações principais, que devem ser consideradas na interpretação dos resultados e que apontam caminhos para investigações futuras. Primeiro, a análise baseou-se em um conjunto restrito de 59 questões, avaliadas por uma única especialista, reduzindo sua generalização e a diversidade de interpretações pedagógicas. Recomenda-se que estudos futuros ampliem o número de itens e envolvam múltiplos avaliadores, a fim de aumentar a robustez analítica e a confiabilidade interavaliador. Segundo, o processo de análise qualitativa demandou cerca de seis horas para apenas 59 itens, indicando alto custo de escalabilidade. Pesquisas futuras devem priorizar o desenvolvimento e a validação de alternativas semi-automatizadas de avaliação pedagógica a fim de manter a relevância da mesma enquanto a torna mais escalável.

Terceiro, não houve um estudo sistemático quanto à formulação dos prompts, motivando investigações futuras que explorarem diferentes estilos, estruturas e níveis de complexidade nos prompts. Por fim, a análise concentrou-se exclusivamente em questões de adição, subtração, multiplicação e divisão, o que não permite conclusões generalizáveis para outros tipos de habilidades matemáticas ou áreas do conhecimento. Recomenda-se que pesquisas futuras ampliem o escopo temático e cognitivo, considerando tarefas mais complexas e diferentes níveis de ensino, para uma avaliação mais abrangente da aplicabilidade de LLMs leves na educação.

Referências

Creswell, J. W. and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Freitas, E., Batista, H. H., Barbosa, G. A., Wenceslau, M., Portela, C., Isotani, S., and Mello, R. F. (2022). Learning analytics desconectada: Um estudo de caso em análise de produções textuais. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 40–49. SBC.

Gasevic, D., Paul, P., Chen, B., Fan, Y., Rodrigo, M., Cobo, C., and Cecilia, A. (2018). Learning analytics for the global south. In *C.P. Lim & V.L. Tinio (Eds.), Foundation*

for Information Technology Education and Development. Foundation for Information Technology Education and Development, Quezon City, Philippines.

Gerganov, G. (2023). llama.cpp.

Graesser, A. C., VanLehn, K., and Rosé, C. P. e. a. (2004). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22(4):39–51.

Guerino, G., Rodrigues, L., Oliveira, L., Marinho, M., Silva, T., Amorim, L., Dermeval, D., da Penha, R., Bittencourt, I., and Isotani, S. (2024). We see you: Understanding math teachers from brazilian public schools to design equitable educational technology. *Revista Brasileira de Informática na Educação*, 32:336–358.

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., and Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, 153:103897.

Indran, I. R., Paranthaman, P., Gupta, N., and Mustafa, N. (2024). Twelve tips to leverage ai for efficient and effective medical question generation: a guide for educators using chat gpt. *Medical Teacher*, 46(8):1021–1026.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (2022). Sistema de avaliação da educação básica 2019: Relatório de desempenho. Technical report, Inep, Brasília, DF.

Isotani, S., Bittencourt, I. I., and Challco, G. C. e. a. (2023). Aied unplugged: Leapfrogging the digital divide to reach the underserved. In *International Conference on Artificial Intelligence in Education*, pages 772–779. Springer.

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., and Kim, H. (2024). Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29(9):11483–11515.

Li, K. and Zhang, Y. (2024). Planning first, question second: An LLM-guided method for controllable question generation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4715–4729, Bangkok, Thailand. Association for Computational Linguistics.

Monteiro Santos, M., Barros, A., and Rodrigues, L. e. a. (2025). Near feasibility, distant practicality: Empirical analysis of deploying and using llms on resource-constrained smartphones. In *Proceedings of the 13th International Conference on Information & Communication Technologies and Development*, ICTD '24, page 224–235, New York, NY, USA. Association for Computing Machinery.

Nikolovski, V., Trajanov, D., and Chorbev, I. (2025). Advancing ai in higher education: A comparative study of large language model-based agents for exam question generation, improvement, and evaluation. *Algorithms*, 18(3).

OECD (2023). Pisa 2022 results (volume i).

Portela, C., Lisbôa, R., Yasojima, K., Cordeiro, T., Silva, A., Dermeval, D., and Isotani, S. (2023). A case study on aied unplugged applied to public policy for learning recovery post-pandemic in brazil. In *International Conference on Artificial Intelligence in Education*, pages 788–796. Springer Nature Switzerland.

Rodrigues, L., Guerino, G., Challco, G. C., Veloso, T. E., Oliveira, L., da Penha, R. S., Melo, R. F., Vieira, T., Marinho, M., Macario, V., et al. (2023). Teacher-centered intelligent tutoring systems: Design considerations from brazilian, public school teachers. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1419–1430. SBC.

Rodrigues, L., Guerino, G., Silva, T. E., Challco, G. C., Oliveira, L., da Penha, R. S., and Isotani, S. (2024a). Mathaide: A qualitative study of teachers' perceptions of an its unplugged for underserved regions. *International Journal of Artificial Intelligence in Education*, pages 1–29.

Rodrigues, L., Guerino, G., Veloso, T. E., Bianchini, L., Xavier, M., Vieira, T., Marinho, M., and Macario, V. (2024b). Mathaide in the classroom: A qualitative analysis of teachers' perspectives of intelligent tutoring systems unplugged. In *Proceedings of the Brazilian Symposium on Computers in Education (SBIE)*. to appear.

Scaria, N., Dharani Chenna, S., and Subramani, D. (2024). Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., and Bittencourt, I. I., editors, *Artificial Intelligence in Education*, pages 165–179, Cham. Springer Nature Switzerland.

Sewunetie, W. and Kovács, L. (2023). Chatgpt in education: Opportunities and concerns for functional english sentence structure analysis for automatic question generation. In *International Conference on Advances of Science and Technology*, pages 79–90. Springer.

Soofi, A. A. and Ahmed, M. U. (2019). A systematic review of domains, techniques, delivery modes and validation methods for intelligent tutoring systems. *International Journal of Advanced Computer Science and Applications*, 10(3).

StatCounter (2024). Mobile vendor market share worldwide. Accessed: 2024-02-14.

Team, G. (2024). Gemma.

Tenório, K., Dermeval, D., Monteiro, M., Peixoto, A., and Silva, A. P. d. (2022). Exploring design concepts to enable teachers to monitor and adapt gamification in adaptive learning systems: a qualitative research approach. *International Journal of Artificial Intelligence in Education*, 32(4):867–891.

Vanlehn, K. (2006). The behavior of tutoring systems. *Int. J. Artif. Intell. Ed.*, 16(3):227–265.

Veloso, T. E., Chalco Challco, G., Rogrigues, L., Versuti, F. M., Sena da Penha, R., Silva Oliveira, L., and Isotani, S. (2023). Its unplugged: Leapfrogging the digital divide for teaching numeracy skills in underserved populations. In *Workshop on International Conference of Artificial Intelligence in Education co-located with The 24th International Conference on Artificial Intelligence in Education*.

Venkatesh, V. and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2):186–204.

Xia, Q., Chiu, T. K., Zhou, X., Chai, C. S., and Cheng, M. (2022). Systematic literature review on opportunities, challenges, and future research recommendations of artifi-

cial intelligence in education. *Computers and Education: Artificial Intelligence*, page 100118.

Xin, C. (2023). Testing llm-based applications: Strategy and challenges. https://blog.scottlogic.com/2023/11/14/testing-LLM-based-applications-strategy-and-challenges.html.

Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., and Irey, R. (2019). The effectiveness of intelligent tutoring systems on k-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, 50(6):3119–3137.