Relato de Experiência

Da engenharia de prompts à mediação pedagógica: concepção e avaliação formativa em tempo real com a plataforma EducAI

From prompt engineering to pedagogical mediation: design and real-time formative assessment with the EducAI platform

Rogério Fabiano Gonçalves [Universidade de Pernambuco, Campus Petrolina | *rogerio.goncalves@upe.br*]
Tarcísio Fulgêncio Alves da Silva [Universidade de Pernambuco, Campus Petrolina | *tarcisio.silva@upe.br*]
Emanuelle Francine Detogni Schmit [Universidade de Pernambuco, Campus Petrolina | *emanuelle.schmit@upe.br*]
Nayara Priscila Dantas de Oliveira [Universidade de Pernambuco, Campus Petrolina | *nayara.oliveira@upe.br*]
Rejane Cristiany Lins de França Pereira [Universidade de Pernambuco, Campus Ouricuri | *rejane.lins@upe.br*]
Paulo Adriano Schwingel [Universidade de Pernambuco, Campus Petrolina | *paulo.schwingel@upe.br*]

Colegiado de Fisioterapia, Universidade de Pernambuco, Campus Petrolina, Rodovia BR 203, Km 2, s/n, Vila Eduardo, Petrolina, Pernambuco, 56328-900, Brasil.

Resumo. A elaboração de avaliações diversificadas, contextualizadas e coerentes com os objetivos de aprendizagem é uma das atividades mais trabalhosas do ensino superior, pois exige itens fiéis ao conteúdo, com distratores plausíveis e diferentes níveis cognitivos. A difusão dos modelos de linguagem de grande porte abriu a possibilidade de automatizar parte dessa tarefa, mas trouxe desafios relativos à fidelidade factual, à qualidade pedagógica dos itens e ao risco de deslocamento do docente de sua função de curadoria. Este artigo objetiva relatar a concepção, a evolução e a aplicação da plataforma EducAI, ferramenta web que utiliza inteligência artificial generativa para produzir questões inéditas e comentadas a partir de materiais didáticos anexados em PDF ou de temas definidos pelo usuário. Trata-se de relato de caráter exploratório, centrado no processo, e não na mensuração de eficácia ou na validação da plataforma. São descritas as decisões de arquitetura: aplicação reunida em contêiner único, com *backend* em *Python* e interface de página única, geração delegada a um modelo acessado por API externa e organização em três perfis de uso — estudante, monitor e docente —, com recursos e permissões distintos. A extração de conteúdo é híbrida, combinando leitura estruturada de documentos nativos e reconhecimento óptico de caracteres para slides e digitalizações; o sistema oferece quatro modos de geração, diferentes formatos de item, estilos de exames brasileiros e associação entre dificuldade e Taxonomia de Bloom revisada, com possibilidade de fixar o nível cognitivo de forma independente. Detalha-se também a evolução da engenharia de prompts, tratada como artefato de software versionável e testável: das primeiras versões, com regras textuais diretas para evitar vícios recorrentes da geração automática, à incorporação de rubricas analíticas para questões abertas e, posteriormente, a uma arquitetura com persona fixa de elaborador de itens, separação entre dificuldade e nível cognitivo, distratores concebidos como diagnósticos e bases de conhecimento estruturadas acionadas de forma seletiva. Complementam esse percurso decisões voltadas à robustez e ao desempenho, como *parser* tolerante a saídas malformadas, mecanismos de autocorreção, balanceamento da posição do gabarito e paralelização de chamadas. A experiência ocorreu entre 2025 e 2026, no âmbito de um edital de inovação pedagógica da Universidade de Pernambuco, em componentes curriculares e na monitoria de um curso da área da saúde, com encontros presenciais integrados às atividades regulares e sessões em que as questões eram projetadas para discussão coletiva antes da apresentação do gabarito comentado. Nos contextos acompanhados, a geração de questões em tempo real integrou-se às aulas como apoio à discussão, enquanto o uso extraclasse permaneceu pontual e variou conforme a preferência dos docentes; como não houve registro sistemático de frequência por modalidade, essa diferença é tratada como aprendizado contextual, e não como comparação de efetividade. Ao longo do desenvolvimento, percepções exploratórias de docentes e estudantes orientaram ajustes sucessivos na ferramenta, apontaram distratores ambíguos e limites na gradação de dificuldade e reforçaram a importância da revisão docente, sem finalidade de mensurar aprendizagem, eficácia ou usabilidade validada. Conclui-se que, nos contextos acompanhados, a geração automatizada de questões adquiriu sentido formativo quando associada à mediação e à validação docente, e que tratar a engenharia de prompts como componente de software sujeito a versionamento e a teste torna as decisões mais rastreáveis, sem garantir, por si só, a qualidade pedagógica.

Abstract. Designing diversified, contextualised assessment items that are consistent with learning objectives is one of the most labour-intensive activities in higher education, as it requires items faithful to the source content, with plausible distractors and varied cognitive levels. The spread of large language models has made it possible to automate part of this task, but it has also raised concerns regarding factual fidelity, the pedagogical quality of the items and the risk of displacing instructors from their curatorial role. This article reports the design, evolution and use of the EducAI platform, a web tool that employs generative artificial intelligence to produce original, annotated assessment items from instructional materials uploaded as PDF files or from topics defined by the user. This is an exploratory report centred on the process rather than on measuring effectiveness or validating the platform. The architectural decisions are described: the application is bundled in a single container, with a Python backend and a single-page interface, generation delegated to a model accessed through an external API, and organisation into three user profiles — student, teaching assistant and instructor —, each with distinct features and permissions. Content extraction is hybrid, combining structured parsing of native documents with optical

character recognition for slides and scanned material; the system offers four generation modes, different item formats, styles drawn from Brazilian national examinations, and an association between difficulty and the revised Bloom's taxonomy, with the option of fixing the cognitive level independently. The article also details the evolution of prompt engineering, treated as a versionable and testable software artefact: from early versions relying on direct textual rules to avoid recurrent flaws of automatic generation, to the incorporation of analytical rubrics for open-ended items and, subsequently, to an architecture combining a fixed item-writer persona, the separation of difficulty from cognitive level, distractors designed as diagnostic, and structured knowledge bases retrieved selectively. This trajectory is complemented by decisions aimed at robustness and performance, such as a parser tolerant of malformed outputs, self-healing mechanisms, balancing of answer-key position and parallelised API calls. The experience took place between 2025 and 2026, within an institutional call for pedagogical innovation at the University of Pernambuco, Brazil, in curricular components and in the tutoring activities of a health sciences programme, with in-person meetings integrated into regular activities and sessions in which items were projected for collective discussion before the annotated answer key was shown. In the observed contexts, real-time item generation was integrated into classes as support for discussion, whereas out-of-class use remained occasional and varied according to instructors' preferences; since no systematic record of frequency by modality was kept, this difference is treated as contextual learning rather than as a comparison of effectiveness. Throughout development, exploratory perceptions from instructors and students guided successive adjustments to the tool, pointed out ambiguous distractors and limitations in difficulty gradation, and reinforced the importance of teacher review, with no intention of measuring learning, effectiveness or validated usability. It is concluded that, in the observed contexts, automated item generation acquired formative meaning when combined with teacher mediation and validation, and that treating prompt engineering as a software component subject to versioning and testing makes decisions more traceable, without in itself guaranteeing pedagogical quality.

Palavras-chave: Inteligência Artificial Generativa, Engenharia de Prompts, Geração Automática de Questões, Avaliação Formativa, Ensino Superior

Keywords: Generative Artificial Intelligence, Prompt Engineering, Automatic Question Generation, Formative Assessment, Higher Education

Recebido/Received: 14 08 2026 • **Aceito/Accepted:** 23 09 2026 • **Publicado/Published:** 15 10 2026

Publicado sob a Licença Pública Internacional Creative Commons Atribuição 4.0 (CC BY 4.0)

1 Introdução

A elaboração de avaliações diversificadas, contextualizadas e cognitivamente exigentes é uma das atividades mais trabalhosas do ensino superior. Construir itens fiéis ao conteúdo, com distratores plausíveis e diferentes níveis cognitivos, consome tempo que poderia ser destinado à mentoria, ao *feedback* e à mediação da aprendizagem. A difusão dos modelos de linguagem de grande porte (*Large Language Models*, LLMs) abriu a possibilidade de automatizar parte dessa tarefa, mas também trouxe desafios: fidelidade factual, qualidade pedagógica dos itens e risco de deslocamento do docente de sua função de curadoria [Zhai and Nehm, 2023; Xia *et al*., 2024].

Estudos recentes têm explorado LLMs para gerar questões a partir de textos, objetivos de aprendizagem e níveis da Taxonomia de Bloom revisada. Os resultados demonstram a viabilidade técnica da geração automática, mas também mostram que a qualidade depende das instruções fornecidas ao modelo e da revisão humana [Lee *et al*., 2024; Scaria *et al*., 2024]. Assim, o desafio não se limita a gerar questões: envolve integrá-las a uma prática avaliativa pedagogicamente orientada.

A plataforma EducAI foi concebida e desenvolvida pelos próprios autores nesse contexto, no âmbito do Edital de Inovação Pedagógica PROGRAD/PFA/UPE nº 05/2025, da Universidade de Pernambuco (UPE). Trata-se de uma aplicação web que emprega inteligência artificial (IA) generativa para criar questões inéditas e comentadas a partir de materiais didáticos anexados em PDF ou de temas especificados pelo usuário. Sua concepção procura conciliar fidelidade ao conteúdo, qualidade pedagógica e possibilidade de uso em tempo real, sem retirar do docente a responsabilidade pela avaliação. Cabe registrar que os autores acumulam as posições de desenvolvedores e de docentes usuários da ferramenta, condição explicitada aqui por seu efeito sobre o alcance do relato: as observações reunidas resultam de um olhar interno ao processo, e não de avaliação independente.

Este artigo objetiva relatar o processo de concepção, evolução e aplicação do EducAI em componentes curriculares da área da saúde. O relato articula três dimensões: as decisões de engenharia de software; a evolução da engenharia de prompts, tratada como artefato versionável e testável; e os aprendizados decorrentes do uso em sala e na monitoria. Trata-se de um relato de caráter exploratório, centrado no processo — concepção, evolução e aplicação — e não na mensuração de eficácia ou na validação da plataforma; seu propósito é sistematizar um percurso de implementação que possa subsidiar iniciativas semelhantes na computação aplicada à educação.

2 Fundamentação Teórica e Trabalhos Relacionados

As metodologias ativas partem do princípio de que a aprendizagem se fortalece quando o estudante ocupa posição central. Tecnologias digitais, quando integradas a um propósito pedagógico, podem favorecer o engajamento e a mobilização de habilidades cognitivas complexas [Freeman *et al*., 2014; Bond *et al*., 2020]. A IA insere-se nessa fronteira como tecnologia potencialmente transformadora, desde que empregada como meio, e não como fim em si [Popenici and Kerr, 2017].

A geração automática de questões (*Automatic Question Generation*, AQG) constitui um campo de pesquisa na interface entre processamento de linguagem natural e educação. Antes da disseminação dos LLMs, os sistemas produziam sobretudo itens de baixo nível cognitivo, frequentemente limitados ao reconhecimento factual [Kurdi *et al*., 2020]. Os LLMs ampliaram essas possibilidades e passaram a produzir itens contextualizados e de níveis cognitivos mais altos, especialmente quando orientados por estratégias de *prompting* alinhadas à Taxonomia de Bloom revisada [Krathwohl, 2002; Scaria *et al*., 2024]. Estratégias com poucos exemplos (*few-shot*) também têm apresentado resultados promissores na geração de questões educacionais [Lee *et al*., 2024].

No campo da avaliação, a IA pode apoiar processos formativos contínuos e oferecer *feedback* oportuno [Zhai and Nehm, 2023; Dönmez, 2024]. Entretanto, revisões recentes destacam problemas de integridade acadêmica, confiabilidade do conteúdo e necessidade de supervisão humana [Xia *et al*., 2024]. No ensino das profissões da saúde, a IA generativa apresenta potencial para apoiar o raciocínio clínico, mas ainda são escassas as evidências robustas de impacto, o que reforça a necessidade de uma integração pedagógica intencional [Pham *et al*., 2025]. A experiência com o EducAI situa-se nessa convergência entre geração automática de itens, avaliação formativa e mediação docente.

3 A Plataforma EducAI

O EducAI é uma plataforma acadêmica de geração e simulação de questões avaliativas, concebida para apoiar professores, monitores e estudantes na elaboração e na prática de avaliações ancoradas em material didático ou em temas indicados pelo usuário.

No plano da arquitetura, a aplicação reúne, em um único contêiner *Docker*, um *backend* em *Python*, desenvolvido com o *framework FastAPI*, e uma interface de página única (SPA) em *React* com *TypeScript*. A interface e a Interface de Programação de Aplicações (API) são servidas na mesma porta, o que simplifica a hospedagem em nuvem e o escalonamento sob demanda. A geração é delegada a um modelo de IA acessado por API externa, da família *DeepSeek*. Registros de uso, avaliações e auditoria são persistidos em planilhas eletrônicas (*Google Sheets*). A sessão do usuário é controlada por token assinado (JWT), que registra o perfil e a área de conhecimento.

A plataforma organiza-se em três perfis. O estudante pratica em modo simulado a partir de conteúdo próprio, com cronômetro opcional e *feedback* configurável. O monitor prepara questões e conduz sessões ao vivo, com projeção do item e opção de revelar ou ocultar o gabarito. O docente cria avaliações, refina questões por IA ou manualmente, audita inconsistências, exporta documentos editáveis e monta provas; as ferramentas de montagem ficam restritas a quem prepara a avaliação (Figura 1).

O usuário pode fornecer arquivos em PDF, como artigos, slides e capítulos digitalizados, ou indicar apenas um tema. A extração é híbrida: documentos nativos são processados por uma biblioteca de extração estruturada, enquanto slides e digitalizações passam por reconhecimento óptico de caracteres (OCR) e pré-processamento de imagem. Em ambos os

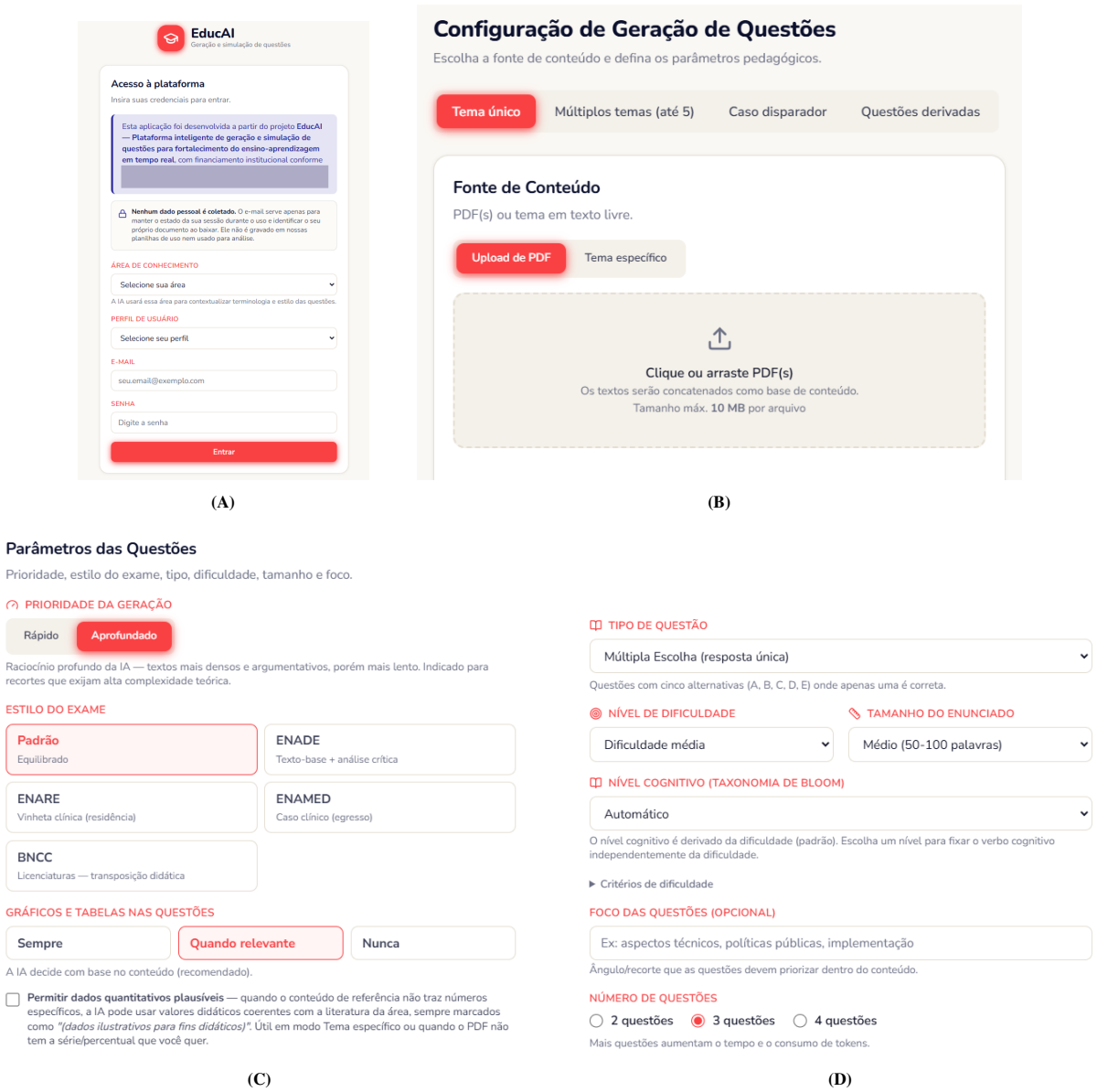


Figura 1. Perfis e configuração do EducAI: (A) tela de acesso, com seleção de área de conhecimento e perfil de usuário; (B) definição do modo de geração e da fonte de conteúdo; (C) parâmetros de geração — prioridade, estilo do exame e uso de gráficos; e (D) definição do tipo, da dificuldade e do nível cognitivo das questões. Fonte: elaboração dos autores (2026).

modos, o prompt orienta o modelo a manter fidelidade ao conteúdo e a não criar fatos, números, autores ou definições ausentes na fonte.

O sistema oferece quatro modos de geração: tema único; múltiplos temas; caso disparador, com cenário-base e questões derivadas; e questões construídas a partir de um item previamente validado. Estão disponíveis questões de múltipla escolha com resposta única, múltipla escolha contextualizada, asserção-razão e dissertativas. O usuário também pode selecionar estilos de exames brasileiros, como ENADE, ENARE e ENAMED. A dificuldade é relacionada à Taxonomia de Bloom revisada, mas o nível cognitivo pode ser definido de forma independente. O sistema ainda produz gráficos, tabelas, *boxplots* e pirâmides etárias, convertidos em imagens nas exportações.

A fidelidade é reforçada por bases de conhecimento estruturadas. Conforme o foco da geração, o sistema acrescenta ao prompt informações de fontes oficiais, como indicadores da Rede Interagencial de Informações para a Saúde (RIPSA) e metas dos Objetivos de Desenvolvimento Sustentável (ODS) [Rede Interagencial de Informações para a Saúde, 2026; Nações Unidas Brasil, 2026]. A recuperação ocorre de forma seletiva, por sobreposição de palavras-chave com limiar mínimo, evitando o acréscimo de contexto sem pertinência temática.

A confiabilidade da saída também exigiu decisões de engenharia de software. Como o modelo devolve texto em formato JSON, um *parser* tolerante recupera a estrutura mesmo diante de saídas malformadas. Mecanismos de autocorreção (*self-healing*) convertem respostas fora do padrão para a estrutura esperada. Um procedimento de balanceamento reduz o viés de posição do gabarito, frequente nos LLMs, e uma persona fixa de elaborador experiente de itens, acompanhada de padrões de qualidade, é utilizada em todas as gerações.

Na aplicação pedagógica, o simulador oferece cronômetro, três níveis de *feedback* e gabarito comentado (Figura 2). A sessão ao vivo projeta o item e mantém, em painel privado,

o gabarito, as explicações e as orientações de estudo. As questões podem ser exportadas em DOCX e HTML.

A plataforma adota medidas de privacidade desde a concepção. As planilhas não armazenam endereços de e-mail em texto claro; o e-mail institucional foi coletado apenas para comunicação e controle da participação; e os registros foram tratados de forma agregada, sem identificação individual. Um Manifesto de Uso Responsável da IA, acessível na interface, explicita que a ferramenta não substitui o docente.

4 Evolução da Engenharia de Prompts

O núcleo intelectual do EducAI é o módulo que constrói os prompts enviados ao modelo. Esse componente foi tratado como artefato de software e evoluiu, de forma iterativa, em resposta às falhas observadas na prática.

Nas primeiras versões, o prompt reunia regras textuais diretas para evitar vícios recorrentes da IA, como alternativas genéricas do tipo "todas as anteriores". A dificuldade era ajustada por instruções básicas, ainda com predomínio de itens de memorização. Também foram observadas questões puramente matemáticas em situações que exigiam análise conceitual. Para corrigir esse desvio, incluiu-se uma regra que subordina a forma da questão ao objetivo pedagógico declarado.

Com a ampliação do uso formativo, o módulo passou a incorporar rubricas analíticas para questões abertas, considerando argumentação, estrutura, fundamentação e aplicação prática. Também foi incluído um mecanismo para adequar o retorno à proficiência informada pelo estudante. O foco, portanto, deslocou-se da simples geração de itens para a produção de *feedback* capaz de apoiar a aprendizagem.

Na etapa seguinte, uma mensagem de sistema estável passou a estabelecer a persona de um elaborador experiente de itens, orientado por princípios de construção de questões (*item-writing*) e pela Taxonomia de Bloom revisada. O nível cognitivo foi separado da dificuldade, e os distratores passaram a ser concebidos como diagnósticos, de modo que cada alternativa incorreta correspondesse a um erro de raciocínio verossímil. Acrescentaram-se padrões para evitar pistas involuntárias, regras para recursos visuais, instruções para efetuar os cálculos quando o foco é quantitativo e as bases de conhecimento descritas na seção anterior.

5 Contexto da Experiência e Sistematização do Relato

A experiência foi desenvolvida no âmbito do Edital de Inovação Pedagógica PROGRAD/PFA/UPE nº 05/2025, do Programa de Fortalecimento Acadêmico da Pró-Reitoria de Graduação da Universidade de Pernambuco, no Campus Petrolina, entre 2025 e 2026. A iniciativa foi articulada aos ODS relativos à educação de qualidade, ao consumo e à produção responsáveis e ao fortalecimento de parcerias [Nações Unidas Brasil, 2026]. A aplicação ocorreu em componentes da graduação em Fisioterapia, com ênfase em Saúde Coletiva: Epidemiologia e Políticas de Saúde, além da monitoria de Epidemiologia.

A seleção dos componentes curriculares decorreu do próprio desenho do edital de inovação pedagógica, que prevê que o docente proponente desenvolva a aplicação nos componentes sob sua responsabilidade, podendo ampliar o escopo a outros componentes cujos docentes manifestem interesse em testar a ferramenta. Não houve, portanto, amostragem probabilística. Os componentes envolvidos pertencem ao curso de Bacharelado em Fisioterapia. A ênfase em Saúde Coletiva — Epidemiologia e Políticas de Saúde — resulta de serem componentes ministrados pelo docente proponente do projeto, condição que permitiu integrar as sessões às atividades regulares sem prejuízo do plano de ensino. Fisioterapia na Atenção Básica e Fisioterapia Musculoesquelética foram incorporados por adesão voluntária de docentes que consentiram em participar: nesses casos, o uso ocorreu sem sessões estruturadas, de modo que não integrou a descrição das aplicações sistematizadas, embora as percepções desses docentes tenham subsidiado o aprimoramento da ferramenta. Docentes de outro curso da área da saúde também utilizaram a plataforma sem vínculo com componente curricular, apenas para conhecê-la e testá-la.

As atividades de 2025 fizeram parte do desenvolvimento da plataforma. O registro das percepções dos participantes ocorreu em 2026, após a aprovação ética. A experiência compreendeu dois períodos: setembro a novembro de 2025, em Epidemiologia, e abril a maio de 2026, em Políticas de Saúde. Foram realizados seis encontros presenciais, três em cada componente curricular, com periodicidade mensal e integrados às atividades regulares. Participaram duas turmas de aproximadamente 20 estudantes cada. Também foram realizadas seis sessões de monitoria de Epidemiologia com os mesmos 20 estudantes dessa turma. Em cada sessão, com duração média de uma hora, eram geradas de três a seis questões.

Quanto aos participantes, não houve seleção individual nem amostragem: as sessões foram conduzidas como atividade coletiva integrada à aula, das quais tomaram parte os estudantes regularmente matriculados, cerca de vinte por turma. As sessões de monitoria foram conduzidas por monitores do próprio componente de Epidemiologia. O registro posterior de percepções foi voluntário e sem contrapartida: quinze estudantes, todos matriculados nos componentes ministrados pelo docente proponente, optaram por registrar suas impressões, assim como oito docentes — cinco do Bacharelado em Fisioterapia e três do Bacharelado em Enfermagem. O critério de inclusão foi ter utilizado a ferramenta durante o semestre letivo; não foram aplicados critérios de exclusão, de modo que foram considerados todos os registros de quem manifestou interesse em opinar sobre a plataforma. Todos os participantes assinaram Termo de Consentimento Livre e Esclarecido.

O módulo interativo era projetado para visualização coletiva. Nas questões de múltipla escolha, os estudantes discutiam as alternativas antes da apresentação do gabarito. Nas questões de cálculo, registravam as respostas em folha própria; nas abertas, debatiam os caminhos possíveis. Após a definição das respostas, o gabarito comentado era exibido e o docente conduzia a discussão.

A sistematização reuniu registros das versões da plataforma, anotações sobre os problemas técnicos e pedagógicos encontrados e a descrição das sessões realizadas. Ao longo desse processo, oito docentes da área da saúde contribuíram



(A) **(B)**

Figura 2. Saídas do EducAI: (A) questão exibida no simulador, com cronômetro; (B) painel com gabarito e justificativas. Fonte: elaboração dos autores (2026).

com percepções exploratórias sobre o percurso de uso, a confiança no conteúdo, a necessidade de edições e eventuais vieses. Quinze estudantes também registraram suas impressões após o uso das questões em debates, monitoria e atividades sobre o Sistema Único de Saúde e as Redes de Atenção à Saúde. Essas percepções subsidiaram os ajustes sucessivos da ferramenta.

Essas contribuições, assim como as observações em sala e os diálogos espontâneos, têm caráter formativo e contextual: orientaram o desenvolvimento da ferramenta até o estágio atual e não constituem investigação sistemática, nem se destinam a estimar eficácia ou a produzir resultados generalizáveis.

Em cumprimento à Resolução nº 466/2012 do Conselho Nacional de Saúde, o projeto foi submetido ao Comitê de Ética em Pesquisa com Seres Humanos do Centro Universitário Integrado de Saúde Amaury de Medeiros (CISAM/UPE) e aprovado em 19 de dezembro de 2025 (CAAE 93622225.9.0000.5191; Parecer nº 8.077.808) [Brasil, Conselho Nacional de Saúde, 2012]. A participação nesta etapa exploratória ocorreu mediante consentimento dos participantes. O projeto de pesquisa dele derivado foi concebido e aprovado pelo Comitê de Ética em Pesquisa; uma coleta de dados abrangente, com instrumentos adequados, constitui etapa posterior, ainda não realizada.

6 Aplicação em Sala e Lições Aprendidas

Nos contextos acompanhados, a aplicação em tempo real mostrou integração mais imediata às aulas e à monitoria. O uso extraclasse ocorreu de forma pontual e variou entre os docentes: alguns se interessaram principalmente pela preparação de avaliações, enquanto outros preferiram explorar a ferramenta fora da sala. Como não houve indicadores sistemáticos de frequência por modalidade, essa diferença deve ser compreendida como aprendizado contextual, e não como comparação de efetividade.

A projeção das questões, seguida de escolha das respostas e debate, converteu a geração automatizada em apoio à mediação pedagógica. As situações produzidas durante a aula permitiram discutir casos não preparados previamente e mobilizar diferentes níveis cognitivos. O valor da atividade, contudo, não decorreu apenas do item gerado, mas da possibilidade de interromper, questionar e contextualizar a resposta com a turma.

O gabarito comentado foi um dos recursos mais valorizados. As justificativas das alternativas ajudaram a esclarecer dúvidas e explicitar equívocos conceituais. Ao mesmo tempo, a mediação do professor permitiu identificar imprecisões, ajustar o vocabulário e discutir por que determinados distratores poderiam induzir interpretações ambíguas.

As percepções reunidas nesta etapa foram coerentes com essa função mediada: a confiança nas questões associou-se à revisão do professor, mais do que ao uso direto da saída automatizada, e os estudantes destacaram os cenários contextualizados, o debate coletivo e as explicações do gabarito. Mais do que avaliar a ferramenta, esses registros ajudaram a identificar o que ajustar e orientaram parte das decisões descritas nas seções anteriores.

Os registros críticos também orientaram mudanças. Alguns distratores foram percebidos como ambíguos; a gradação de dificuldade nem sempre permitiu ao estudante reconhecer seu nível; e um dos cenários sobre Redes de Atenção à Saúde concentrou-se em vocabulário técnico, sem exigir análise suficiente da organização dos serviços. Esses registros reforçaram a necessidade de revisão docente e de testes contínuos dos prompts.

Por definição, este relato não comporta um desenho destinado a medir aprendizagem, desempenho ou mudança na prática docente. As percepções reunidas tiveram finalidade exploratória e informaram o desenvolvimento, sem constituir evidência de eficácia ou de usabilidade validada; tampouco

houve medida sistemática da adesão a cada modalidade. O alcance do artigo restringe-se, portanto, à descrição crítica do processo de concepção, evolução e aplicação da plataforma e aos aprendizados que dele decorreram.

7 Desafios e Soluções

A implementação reuniu desafios relacionados à confiabilidade do conteúdo, à estabilidade da saída, ao desempenho, à recuperação de informações e à adoção pelos usuários.

O risco central de sistemas baseados em LLM é a produção de conteúdo fluente, mas incorreto [Sahoo *et al.*, 2024]. A resposta combinou regras de fidelidade no prompt, ancoragem em fontes oficiais e manutenção do docente como revisor final. Mudanças entre gerações de modelos podem melhorar a adesão às instruções, mas também alteram o comportamento da ferramenta; por isso, a atualização do modelo exige novos testes e não elimina a validação humana.

Os testes iniciais e as avaliações docentes identificaram inconsistências na geração. Foram adotados refinamento iterativo dos prompts, auditoria com registro do estado original e do editado, *parser* tolerante e mecanismos de autocorreção. O viés de posição do gabarito, que concentrava respostas corretas em determinadas letras, foi tratado por balanceamento em lote, preservando o conteúdo e a alternativa correta de cada item.

O tempo de resposta foi um gargalo para o uso em sala. Em algumas configurações, a geração podia levar vários minutos. Duas decisões reduziram essa espera: priorizar respostas mais rápidas, sem acionar por padrão modos ampliados de raciocínio (*thinking*), e paralelizar as chamadas de API em solicitações mais complexas. Em teste interno com três questões, o tempo médio observado passou de aproximadamente 150 para 24 segundos. Trata-se de um indicador técnico de processo, condicionado à configuração utilizada, e não de uma avaliação formal de desempenho.

Parte das falhas inicialmente atribuídas ao LLM decorria da recuperação da informação ou da qualidade dos arquivos de entrada. Foram observadas dificuldades na leitura de um PDF e na extração de dados quantitativos de tabelas. Para enfrentá-las, aprimorou-se o fluxo híbrido de extração, com OCR e pré-processamento de imagens, e implementou-se uma estratégia simplificada de recuperação aumentada por geração (RAG). Documentos extensos são segmentados em trechos (*chunks*), classificados por relevância temática, e apenas os segmentos mais pertinentes compõem o contexto enviado ao modelo.

Quanto à adoção, as preferências dos docentes não foram uniformes. O uso extraclasse mostrou utilidade para exploração e preparação de avaliações, mas sua incorporação à rotina permaneceu ocasional no período acompanhado. O uso em tempo real apresentou encaixe mais direto nas atividades conduzidas pelos autores. As duas modalidades, portanto, devem ser tratadas como complementares, e não como alternativas concorrentes.

Na infraestrutura, a geração é limitada principalmente pela espera de rede. Uma hospedagem econômica, com hibernação sob demanda, atende ao uso acadêmico esporádico. Em situações de maior demanda, como períodos de prova, prevê-se escalonamento temporário de memória e réplicas, seguido de retorno à configuração regular para controle dos custos.

8 Considerações Finais e Perspectivas

A principal contribuição da experiência com o EducAI não está em utilizar inteligência artificial para gerar questões, mas em mostrar como essa geração pode ser integrada à avaliação formativa sem retirar do docente a responsabilidade pelas decisões pedagógicas. A plataforma atuou como suporte dinâmico à discussão, e sua evolução técnica respondeu a necessidades observadas durante o uso.

Nos contextos acompanhados, a aplicação em tempo real apresentou integração mais direta às aulas e à monitoria, enquanto o uso extraclasse dependeu das preferências e das rotinas dos docentes. Essa observação não estabelece superioridade entre modalidades. Indica, antes, que o valor da ferramenta depende de sua adequação à atividade pedagógica e da presença de mecanismos de revisão, contextualização e *feedback*.

Os próximos ciclos concentram-se no amadurecimento da plataforma: o aprimoramento do uso extraclasse, com armazenamento de questões favoritas, simulado programado e painel de desempenho da turma; a expansão para outros campos, incluindo licenciaturas e recursos relacionados à Base Nacional Comum Curricular [Brasil, Ministério da Educação, 2018]; e a produção de documentação que favoreça a transferência tecnológica.

Do ponto de vista computacional, a experiência mostrou a utilidade de tratar a engenharia de prompts como componente de software sujeito a versionamento, testes e aprimoramento contínuo. Essa abordagem não garante, por si só, a qualidade pedagógica, mas torna as decisões mais rastreáveis e facilita a correção de falhas observadas na prática.

Declarações complementares

Agradecimentos

Os autores agradecem à Pró-Reitoria de Graduação da Universidade de Pernambuco (PROGRAD/UPE), pelo apoio concedido por meio do Programa de Fortalecimento Acadêmico (PFA); aos monitores do componente curricular de Epidemiologia, pela condução das sessões de monitoria; e aos docentes e estudantes que participaram das atividades e compartilharam suas percepções sobre a plataforma.

Financiamento

Esta experiência foi desenvolvida com apoio do Edital de Inovação Pedagógica PROGRAD/PFA/UPE nº 05/2025, do Programa de Fortalecimento Acadêmico (PFA) da Pró-Reitoria de Graduação da Universidade de Pernambuco (UPE), que também concedeu bolsa BIA vinculada ao projeto.

Contribuições dos autores

RFG é o principal contribuidor e escritor deste manuscrito e participou de todas as etapas do trabalho, incluindo administração do projeto e supervisão. Todos os autores contribuíram para a conceituação do estudo. RFG foi responsável pelo desenvolvimento do software. RFG, NPDO, EFDS e RCLFP realizaram a investigação. RFG redigiu o rascunho original; RFG, TFAS, EFDS e NPDO participaram da escrita – revisão e edição. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não há conflitos de interesse.

Disponibilidade de artefatos de pesquisa

Os dados e outros materiais criados e/ou usados neste relato serão disponibilizados mediante solicitação aos autores.

Uso de Inteligência Artificial

As ferramentas ChatGPT (OpenAI; modelos da linha Pro, com raciocínio estendido) e Claude (Anthropic; modelos de nível avançado, com raciocínio estendido), nas versões disponibilizadas entre julho e setembro de 2026, foram utilizadas para apoiar a revisão gramatical e estilística, o aprimoramento da clareza e a organização textual deste manuscrito, bem como a formatação do texto no modelo LaTeX do periódico. Não foram empregadas na coleta ou na análise dos dados. Todas as sugestões incorporadas foram criticamente revisadas e validadas pelos autores, que assumem integral responsabilidade pelo conteúdo final. Não foi atribuída autoria às ferramentas de inteligência artificial. Registra-se, adicionalmente, que a plataforma relatada neste artigo emprega um modelo de linguagem de grande porte da família DeepSeek como componente do objeto de estudo, e não como recurso de escrita do manuscrito.

Outras informações relevantes

Em cumprimento à Resolução nº 466/2012 do Conselho Nacional de Saúde, o projeto foi submetido ao Comitê de Ética em Pesquisa com Seres Humanos do Centro Universitário Integrado de Saúde Amaury de Medeiros (CISAM/UPE) e aprovado em 19 de dezembro de 2025 (CAAE 93622225.9.0000.5191; Parecer nº 8.077.808). A participação de docentes e estudantes ocorreu mediante assinatura de Termo de Consentimento Livre e Esclarecido. Nenhum dado pessoal identificável foi armazenado ou analisado.

Referências

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., and Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: a systematic evidence map. *International Journal of Educational Technology in Higher Education*, 17(1):2. DOI: 10.1186/s41239-019-0176-8.

Brasil, Conselho Nacional de Saúde (2012). Resolução nº 466, de 12 de dezembro de 2012. `https://conselho.saude.gov.br/resolucoes/2012/Reso466.pdf`.

Brasil, Ministério da Educação (2018). Base nacional comum curricular. `https://cdn.mec.gov.br/basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf`. Acesso em: 27 jul. 2026.

Dönmez, M. (2024). AI-based feedback tools in education: a comprehensive bibliometric analysis study. *International Journal of Assessment Tools in Education*, 11(4):622–646. DOI: 10.21449/ijate.1467476.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415. DOI: 10.1073/pnas.1319030111.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Practice*, 41(4):212–218. DOI: 10.1207/s15430421tip4104_2.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204. DOI: 10.1007/s40593-019-00186-y.

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. DOI: 10.1007/s10639-023-12249-8.

Nações Unidas Brasil (2026). Objetivos de desenvolvimento sustentável. `https://brasil.un.org/pt-br/sdgs`. Acesso em: 29 jul. 2026.

Pham, T. D., Karunaratne, N., Exintaris, B., Liu, D., Lay, T., Yuriev, E., and Lim, A. (2025). The impact of generative AI on health professional education: a systematic review in the context of student learning. *Medical Education*, 59(12):1280–1289. DOI: 10.1111/medu.15746.

Popenici, S. A. D. and Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(22). DOI: 10.1186/s41039-017-0062-8.

Rede Interagencial de Informações para a Saúde (2026). *Indicadores básicos para a saúde no Brasil: conceitos e aplicações*. Organização Pan-Americana da Saúde, Brasília, DF, 3 edition. Disponível em: `https://www.ripsa.org.br/publicacoes-ripsa/`. Acesso em: 29 jul. 2026.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: techniques and applications. *arXiv preprint arXiv:2402.07927*. DOI: 10.48550/arXiv.2402.07927.

Scaria, N., Chenna, S. D., and Subramani, D. (2024). Automated educational question generation at different Bloom's skill levels using large language models: strategies and evaluation. In *Artificial Intelligence in Education (AIED 2024)*, volume 14830 of *Lecture Notes in Computer Science*, pages 165–179. Springer, Cham. DOI: 10.48550/arXiv.2408.04394.

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., and Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, 21(1):40. DOI: 10.1186/s41239-024-00468-z.

Zhai, X. and Nehm, R. H. (2023). AI and formative assessment: the train has left the station. *Journal of Research in Science Teaching*, 60(6):1390–1398. DOI: 10.1002/tea.21885.