

# Uma Proposta de Sistema Tutor Inteligente para a Disciplina Cálculo Diferencial

Pedro Gabriel de Carvalho Cruz<sup>1</sup>, Anderson Pinheiro Cavalcanti<sup>1</sup>,  
Roy Percy Tocto Guarniz<sup>1</sup>, Eber Daniel Chuño Vizarreta<sup>1</sup>

<sup>1</sup> Universidade Federal Rural de Pernambuco (UFRPE)  
Unidade Acadêmica de Belo Jardim (UABJ)

pedro.carvalhocruz@ufrpe.br

**Resumo.** Disciplinas de Cálculo Diferencial impõem elevada carga conceitual e demandam acompanhamento docente contínuo. Este artigo apresenta o Solve With Me, um Sistema Tutor Inteligente que integra geração automática de questões por IA generativa, resolução de atividades de múltipla escolha e agrupamento de perfis de aprendizagem por K-Means. Trinta questões geradas por dois modelos foram avaliadas por dois professores quanto à corretude, clareza e relevância, obtendo boa aceitação (médias de 4,03 e 3,93 em 5) e concordância substancial sobre a aprovação; a corretude matemática foi o principal ponto de atenção. Dados fictícios de uso ilustram a viabilidade técnica do agrupamento para apoiar decisões pedagógicas. Espera-se que a ferramenta contribua para o aprimoramento do ensino de Cálculo e para a redução de evasão e reprovação.

**Abstract.** Differential Calculus courses impose a high conceptual load and demand continuous teacher support. This paper presents Solve With Me, an Intelligent Tutoring System that integrates AI-based automatic question generation, multiple-choice problem solving, and K-Means clustering of learning profiles. Thirty AI-generated questions were assessed by two teachers for correctness, clarity, and relevance, achieving good acceptance (averages of 4.03 and 3.93 out of 5) and substantial agreement on approval; mathematical correctness was the main concern. Fictitious usage data illustrate the technical feasibility of clustering to support pedagogical decisions. The tool is expected to contribute to improving the teaching of Calculus and to reducing dropout and failure rates.

## 1. Introdução

A aprendizagem de matemática permanece como um dos principais desafios do cenário educacional brasileiro, especialmente em disciplinas caracterizadas por elevado nível de abstração e formalismo matemático, como o Cálculo Diferencial e Integral (CDI). Dados do Programa Internacional de Avaliação de Estudantes (PISA) de 2022 indicam que aproximadamente 73% dos estudantes brasileiros apresentaram desempenho insuficiente em matemática [BRASIL 2023], evidenciando lacunas formativas que se refletem diretamente no ingresso e permanência de estudantes em cursos das áreas de Ciências Exatas e Engenharias. Essas dificuldades tornam-se ainda mais evidentes no ensino superior, onde disciplinas como Cálculo Diferencial e Integral, Geometria Analítica e Álgebra Linear frequentemente apresentam elevados índices de reprovação e evasão [SANTOS et al. 2024].

Ao mesmo tempo, tecnologias digitais e recursos baseados em Inteligência Artificial (IA) têm ampliado as possibilidades de personalização do ensino, acompanhamento de estudantes e apoio ao trabalho docente [Woolf 2009, Baker and Inventado 2014, Kasneci et al. 2023a]. Em particular, Sistemas Tutores Inteligentes (STIs) buscam oferecer suporte adaptativo ao estudante, registrando interações, identificando dificuldades e disponibilizando atividades compatíveis com seu estado de aprendizagem [Woolf 2009, VanLehn 2011]. Essa abordagem é especialmente relevante em disciplinas de Cálculo, nas quais erros recorrentes podem estar associados a lacunas específicas, como manipulação algébrica, interpretação de enunciados, escolha de métodos de resolução ou aplicação incorreta de propriedades matemáticas [Tall 1993].

Nesse cenário, modelos de IA generativa também vêm sendo explorados para apoiar a elaboração de materiais didáticos e avaliações [Kasneci et al. 2023a, Yan et al. 2024]. A geração automática de questões pode reduzir parte do esforço docente necessário para construir bancos de itens, além de permitir a criação de exercícios parametrizados por assunto, dificuldade e competência [Kurdi et al. 2020b]. Contudo, em domínios formais como a matemática, a geração automática exige validação rigorosa, pois uma questão aparentemente clara pode apresentar erro conceitual, alternativa incorreta, distratores frágeis ou explicação inadequada [Lee et al. 2024a, Feng et al. 2024b, Grévisse et al. 2024]. Assim, a integração entre IA generativa e revisão docente torna-se fundamental para garantir qualidade pedagógica e matemática [Lee et al. 2024a, Grévisse et al. 2024].

Este trabalho apresenta o *Solve With Me*, uma ferramenta educacional concebida como um Sistema Tutor Inteligente para apoiar professores e estudantes em disciplinas de Cálculo. A ferramenta permite que professores gerem questões automaticamente a partir de um tópico da disciplina e que estudantes resolvam questões de múltipla escolha no ambiente. A partir dos registros de acertos, erros e interações, o sistema calcula indicadores de desempenho e utiliza o algoritmo K-Means para agrupar estudantes com perfis semelhantes, oferecendo subsídios para personalização do ensino e tomada de decisão pedagógica. As principais contribuições deste trabalho são: (i) a apresentação de uma ferramenta web que integra geração automática de questões e resolução de atividades por estudantes; (ii) uma avaliação preliminar de questões geradas por IA com dois professores da área de Cálculo; (iii) uma análise técnica de um banco de dados fictício utilizado para validar o registro de interações; e (iv) uma proposta de uso de agrupamento por K-Means para apoiar a identificação de perfis de aprendizagem.

## 2. Trabalhos Relacionados

Os Sistemas Tutores Inteligentes formam uma linha clássica de pesquisa em IA aplicada à educação, com origem em sistemas de instrução assistida por computador capazes de adaptar o processo de ensino às necessidades do estudante [Carbonell 1970]. Em geral, STIs combinam modelos de domínio, modelos de estudante, estratégias de tutoria e interfaces de interação. No contexto de disciplinas matemáticas, esses sistemas podem ser utilizados para registrar passos de resolução, identificar erros recorrentes e oferecer feedback ou atividades direcionadas. No *Solve With Me*, essa perspectiva é incorporada por meio do registro de tentativas, respostas e desempenho por tópico, que alimentam o modelo de agrupamento dos estudantes.

A geração automática de questões também possui trajetória consolidada na literatura educacional. Revisões sistemáticas apontam que a área busca reduzir o custo de elaboração manual de itens e ampliar a disponibilidade de exercícios para avaliação formativa [Kurdi et al. 2020a]. Com o avanço dos grandes modelos de linguagem (LLMs - Large Language Models), novas abordagens passaram a explorar LLMs para gerar enunciados, alternativas, distratores e explicações [Wang et al. 2025, Cheng et al. 2025]. Entretanto, estudos recentes indicam que a geração de questões matemáticas de múltipla escolha ainda apresenta desafios, especialmente na criação de distratores plausíveis e na garantia de corretude matemática [Scarlatos et al. 2024a]. Em português, trabalhos recentes também vêm investigando a geração e avaliação de itens educacionais com LLMs [Santos et al. 2025, Nogueira et al. 2025, Quincozes et al. 2025].

Por fim, a mineração de dados educacionais fornece métodos para analisar registros de interação em ambientes digitais de aprendizagem [Ferreira-Mello et al. 2019]. Técnicas de agrupamento, como o K-Means, podem ser utilizadas para identificar perfis de estudantes com base em padrões de desempenho, engajamento e dificuldade; permitindo observar grupos de estudantes que apresentam necessidades semelhantes, como baixo desempenho em determinado tópico, alta atividade com baixa taxa de acerto ou desempenho consistente em diferentes conteúdos.

### 3. Método

Este estudo apresenta o desenvolvimento e a avaliação preliminar do *Solve With Me*, uma ferramenta web de apoio ao ensino-aprendizagem em disciplinas de Cálculo. A metodologia foi organizada em quatro etapas: desenvolvimento do protótipo, geração automática de questões, avaliação docente das questões geradas e análise técnica dos dados de interação.

#### 3.1. Visão geral da ferramenta

O *Solve With Me* é uma plataforma web educacional desenvolvida para apoiar professores e estudantes em atividades de resolução de problemas matemáticos. A proposta do sistema é registrar as interações dos estudantes durante a resolução de questões e transformar esses registros em indicadores pedagógicos, permitindo que o professor acompanhe dificuldades individuais e coletivas da turma. Essa concepção aproxima-se dos Sistemas Tutores Inteligentes, que utilizam registros de interação, modelos de estudante e estratégias de tutoria para apoiar o processo de aprendizagem [Woolf 2009, VanLehn 2011].



Figura 1. Fluxograma de utilização da ferramenta.

A plataforma está organizada em dois perfis principais: professor/administrador e aluno. No perfil do professor, é possível criar turmas, gerar códigos de acesso, cadastrar

disciplinas, tópicos e questões, além de acompanhar indicadores de desempenho. As questões podem conter enunciado textual, imagens, expressões matemáticas em  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  e uma ou mais perguntas objetivas. O fluxo de utilização da ferramenta pelo professor é resumido na Figura 1. Inicialmente, o professor precisa criar a turma, gerando um código que permite o ingresso dos alunos. O professor também pode adicionar outro professor ou monitor da disciplina com perfil de administrador, para cadastrar assuntos e questões e acompanhar a turma ao longo do semestre. No perfil do estudante, o sistema disponibiliza atividades de múltipla escolha, registra as respostas e calcula indicadores de desempenho por tópico. Esses registros são armazenados em banco de dados e utilizados posteriormente para construção de perfis de aprendizagem.

### 3.2. Geração automática de questões

O módulo de geração automática recebe como entrada informações como disciplina, tópico, nível de dificuldade e tipo de questão. A partir desses parâmetros, um modelo de IA generativa produz o enunciado, alternativas, alternativa correta e explicação da resolução. Como a geração automática pode produzir erros matemáticos ou explicações inadequadas, as questões não são consideradas automaticamente válidas. Embora modelos generativos possam auxiliar na criação de materiais didáticos, questões matemáticas precisam de validação rigorosa para garantir a corretude conceitual, clareza do enunciado, consistência das alternativas e plausibilidade dos distratores [Kurdi et al. 2020c, Kasneci et al. 2023b, Lee et al. 2024b, Feng et al. 2024a, Scarlatos et al. 2024b]. Elas devem passar por avaliação docente, especialmente quanto à corretude matemática, clareza do enunciado e relevância para o conteúdo. A Figura 2 apresenta a interface de geração, na qual o professor define os parâmetros e fornece orientações para a IA, com pré-visualização do enunciado, do  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  e das alternativas antes de salvar a questão.

**Figura 2.** Interface de geração assistida por IA, com definição de parâmetros, orientações ao modelo e pré-visualização da questão antes da revisão docente.

### 3.3. Avaliação das questões por professores

Para avaliar preliminarmente a qualidade das questões geradas, foram produzidas 30 questões por dois modelos de IA generativa: o *Gemini-2.5-flash* e o modelo *GPT-5.4-nano*. As questões abrangeram três tópicos da disciplina (limite por fatoração, regra da cadeia e integral dupla), com cinco questões por tópico para cada modelo, totalizando 15

itens por modelo e 30 itens no conjunto. A geração foi parametrizada por um mesmo *prompt* estruturado, que solicitava questões de múltipla escolha com cinco alternativas, exatamente uma correta, distratores baseados em erros típicos de estudantes (troca de sinal, limites de integração invertidos, aplicação incorreta da regra da cadeia, entre outros), expressões em  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  e explicação passo a passo, além de metadados de nível de dificuldade (médio/difícil) e nível taxonômico de Bloom.

As questões foram avaliadas de forma independente por dois professores da área de Cálculo. Cada avaliação considerou três critérios: corretude matemática (em três níveis: correta, parcialmente correta ou incorreta), clareza do enunciado e relevância do conteúdo (ambos em escala Likert de 1 a 5). Além disso, os avaliadores atribuíram uma nota geral em escala de 1 a 5 e classificaram a questão como aprovada ou rejeitada. Para fins de análise comparativa, os rótulos categóricos da corretude foram mapeados para a mesma escala numérica de 1 a 5 (incorreta = 1, parcialmente correta = 3, correta = 5).

As análises quantitativas incluíram médias, medianas, taxas de aprovação, distribuição dos critérios e medidas de concordância entre revisores. Para a concordância, foram calculadas a concordância exata, o Kappa de Cohen, o Kappa ponderado quadrático para escalas ordinais e o coeficiente de correlação intraclasse. Os comentários textuais dos revisores foram analisados de forma exploratória para identificar problemas recorrentes.

### 3.4. Agrupamento dos estudantes por K-Means

Para testar a ferramenta, foi utilizado um banco SQLite fictício contendo registros de turmas, disciplinas, tópicos, desafios, perguntas, tentativas e interações. A base foi utilizada para validar a persistência das interações e demonstrar a extração de indicadores para agrupamento. Por se tratar de uma base fictícia, os resultados derivados desse banco não são interpretados como evidência de aprendizagem real, mas como validação técnica do fluxo de dados.

O módulo de personalização utiliza registros de interação para agrupar estudantes com perfis semelhantes. A partir do banco de dados, são extraídas variáveis como número de tentativas, número de interações, taxa geral de acerto, taxa de acerto por tópico e quantidade de tópicos acessados. Essas variáveis são padronizadas antes da aplicação do K-Means, uma vez que possuem escalas diferentes. O número de grupos pode ser definido por critérios exploratórios, como análise do cotovelo e coeficiente de silhueta, ou por necessidade pedagógica do professor. Após a execução do algoritmo, os grupos são interpretados pedagogicamente a partir de seus indicadores médios. Por exemplo, um grupo pode representar estudantes com baixa atividade e baixo desempenho, enquanto outro pode representar estudantes com alta atividade, mas dificuldades persistentes em determinado tópico.

## 4. Resultados

### 4.1. Funcionalidades do protótipo

O protótipo permite que professores cadastrem conteúdos e gerem questões automaticamente a partir de tópicos da disciplina. As questões geradas podem ser revisadas antes de serem disponibilizadas aos estudantes. No ambiente do estudante, as atividades são apresentadas em formato de múltipla escolha, e cada resposta recebe feedback imediato, indicando acerto ou erro e a alternativa correta (Figura 3). Cada resposta é registrada pelo

sistema. Esses registros permitem calcular indicadores como taxa de acerto, número de tentativas e desempenho por tópico, que posteriormente alimentam o agrupamento dos estudantes.

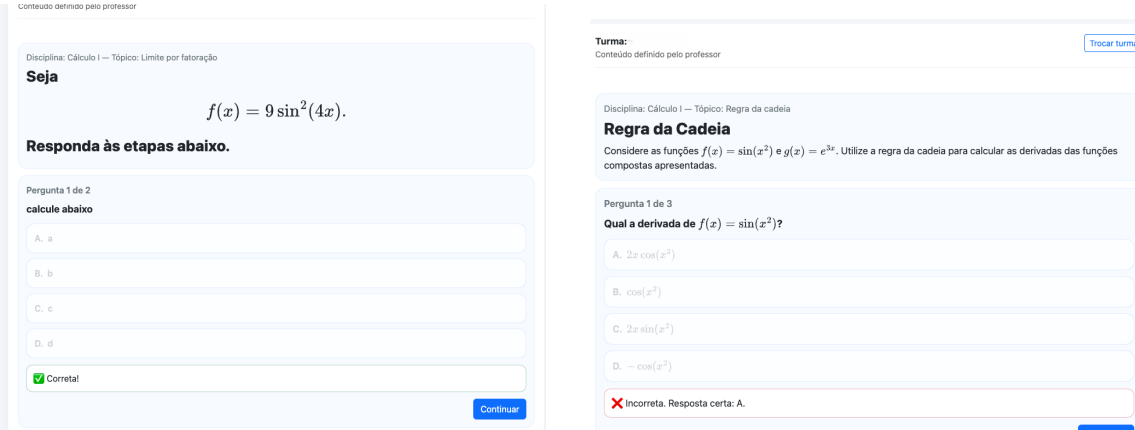


Figura 3. Ambiente do estudante: resolução de questões de múltipla escolha com feedback imediato de acerto (esquerda) e de erro com indicação da alternativa correta (direita).

A partir desses registros, o painel docente organiza uma fila de acompanhamento que ordena os estudantes por prioridade, combinando indicadores de risco, taxa de erro, engajamento e tópicos mais sensíveis (Figura 4). Esse recurso busca apoiar o professor na decisão sobre por onde iniciar o acompanhamento da turma.

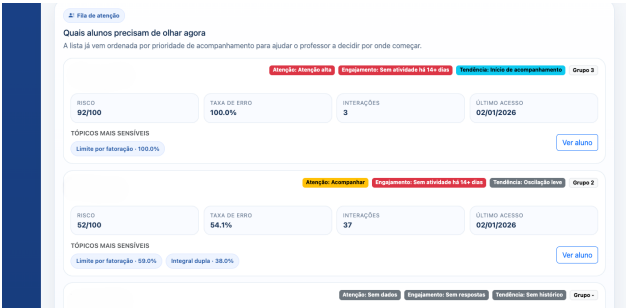


Figura 4. Fila de acompanhamento que ordena os estudantes por prioridade a partir de indicadores de risco, engajamento e desempenho por tópico.

#### 4.2. Avaliação das questões geradas

Foram analisadas 60 avaliações correspondentes a 30 questões únicas, avaliadas por dois professores. O Revisor 1 atribuiu score médio de 4,03, enquanto o Revisor 2 atribuiu score médio de 3,93. As taxas de aprovação foram de 63,3% e 73,3%, respectivamente. Quando comparados os modelos de geração, as questões produzidas pelo modelo GPT-5.4-nano apresentaram score médio de 4,20 para ambos os revisores, enquanto as questões geradas pelo Gemini-2.5-flash obtiveram médias de 3,87 e 3,67 (média geral de 4,20 para o GPT-5.4-nano e 3,77 para o Gemini-2.5-flash). A Tabela 1 consolida os resultados por critério e por revisor.

Quanto aos critérios específicos, a maior média foi observada em relevância do conteúdo (4,10), seguida por clareza do enunciado (4,05). O critério com menor média

foi corretude matemática (3,63), indicando que a principal limitação das questões geradas está relacionada à precisão conceitual e à validade matemática das resoluções e alternativas. Na distribuição desse critério, 48,3% das avaliações classificaram as questões como corretas, 35,0% como parcialmente corretas e 16,7% como incorretas.

**Tabela 1. Resultados da avaliação docente, por revisor e no conjunto (escala de 1 a 5).**

| Métrica / Critério     | Revisor 1 | Revisor 2 | Geral |
|------------------------|-----------|-----------|-------|
| Score geral            | 4,03      | 3,93      | 3,98  |
| Taxa de aprovação      | 63,3%     | 73,3%     | 68,3% |
| Corretude matemática   | 3,73      | 3,53      | 3,63  |
| Clareza do enunciado   | 4,10      | 4,00      | 4,05  |
| Relevância do conteúdo | 4,17      | 4,03      | 4,10  |

A concordância entre revisores foi alta para o status final de aprovação/rejeição, com concordância exata de 90% e Kappa de 0,772. Para o score ordinal, a concordância exata foi de 50%, mas 93,3% das questões apresentaram diferença de até um ponto entre os revisores. O Kappa ponderado quadrático para o score foi 0,728, e o coeficiente de correlação intraclass foi 0,734. Esses resultados sugerem que, embora existam diferenças pontuais na atribuição das notas, os avaliadores apresentaram concordância consistente sobre a qualidade geral dos itens.

#### 4.3. Análise técnica do banco de dados

O banco fictício contém registros de três tópicos: limite por fatoração, regra da cadeia e integral dupla. Foram identificadas 70 tentativas de desafios e 108 interações. A maior parte das interações ocorreu no tópico limite por fatoração, com 75 interações e taxa de acerto de 37,3%. O tópico integral dupla apresentou 28 interações e taxa de acerto de 60,7%. Já regra da cadeia apresentou apenas 5 interações, com taxa de acerto de 20,0%.

**Tabela 2. Perfis de estudantes obtidos pelo K-Means ( $k = 3$ ) sobre a base fictícia (médias por grupo).**

| Grupo | <i>n</i> | Tent. | Inter. | Acerto | Interpretação                            |
|-------|----------|-------|--------|--------|------------------------------------------|
| G1    | 4        | 3,3   | 5,8    | 72%    | Baixa atividade, bom desempenho          |
| G2    | 1        | 44,0  | 62,0   | 40%    | Alta atividade, desempenho intermediário |
| G3    | 4        | 3,3   | 5,8    | 24%    | Baixa atividade, baixo desempenho        |

A base permitiu demonstrar a extração de indicadores por estudante e a aplicação do K-Means. Dos 12 estudantes matriculados, 9 possuíam interações registradas e foram considerados no agrupamento, descrito por quatro variáveis padronizadas: número de tentativas, número de interações, taxa geral de acerto e número de tópicos acessados. Em uma solução exploratória com três grupos (coeficiente de silhueta de 0,32), foram identificados perfis interpretáveis, resumidos na Tabela 2: um grupo com baixa atividade e bom desempenho, um único estudante com atividade muito alta e desempenho intermediário, e um grupo com baixa atividade e baixo desempenho. A Figura 5 ilustra a visão individual gerada pela ferramenta, que reúne o índice de risco, a relação entre acertos e erros, os tópicos críticos, os melhores sinais e ações sugeridas, além da visão de perfis da turma derivada do agrupamento. Entretanto, como o banco é fictício e apresenta inconsistências referenciais, esses agrupamentos devem ser interpretados apenas como validação técnica do pipeline.



**Figura 5. Visão individual do estudante com indicadores de risco, distribuição de acertos e erros, tópicos críticos, ações sugeridas e perfis da turma derivados do agrupamento por K-Means.**

## 5. Considerações Finais

Este artigo apresentou o *Solve With Me*, uma proposta de Sistema Tutor Inteligente que integra geração automática de questões com IA e agrupamento de estudantes por K-Means em disciplinas de Cálculo. A ferramenta busca apoiar professores na criação de atividades e oferecer subsídios para personalização do ensino a partir dos registros de interação dos estudantes. A avaliação preliminar de 30 questões geradas por IA indicou boa aceitação geral pelos revisores, com alta concordância quanto ao status de aprovação/rejeição. No entanto, a corretude matemática emergiu como o principal desafio, reforçando a necessidade de revisão docente antes do uso pedagógico das questões.

O uso do K-Means mostrou-se tecnicamente viável para agrupar estudantes a partir de indicadores de interação. Entretanto, a interpretação pedagógica dos grupos depende da qualidade dos dados coletados e da escolha adequada das variáveis. Além disso, agrupamentos não devem ser usados de forma determinística para rotular estudantes, mas como instrumentos de apoio à decisão docente. O professor deve interpretar os perfis em conjunto com evidências qualitativas, histórico de participação e contexto da turma.

As limitações deste estudo incluem o uso de uma base fictícia para simular interações de estudantes, o número reduzido de avaliadores docentes e a ausência de uma avaliação longitudinal com estudantes reais. Como trabalho em andamento, os resultados devem ser interpretados como evidências preliminares sobre a viabilidade da ferramenta e como base para estudos futuros em turmas reais.

Como trabalhos futuros, pretende-se aprimorar os prompts de geração de questões, incluir verificações automáticas de formatação e consistência matemática, ampliar a avaliação docente dos itens e conduzir estudos com estudantes reais para investigar o impacto da ferramenta no desempenho, engajamento e acompanhamento pedagógico em disciplinas de Cálculo.

## Agradecimentos

À Pró-Reitoria de Extensão, Cultura e Cidadania da Universidade Federal Rural de Pernambuco (PROExC/UFRPE), pela bolsa e ajuda de custo para a execução do projeto.

## Declaração sobre uso de Inteligência Artificial

Reconhecemos o uso de IA Generativa para auxiliar na elaboração e revisão deste artigo, que foi devidamente revisado, e assumimos total responsabilidade por seu conteúdo.

## Cuidados Éticos

A avaliação envolveu apenas a apreciação técnica de questões geradas automaticamente por dois professores especialistas da área, atuando como revisores dos artefatos produzidos, sem coleta de dados pessoais sensíveis nem intervenção em seres humanos. Os dados de uso da plataforma utilizados neste estudo são fictícios e não correspondem a estudantes reais. Estudos futuros com estudantes reais serão submetidos a apreciação de Comitê de Ética em Pesquisa.

## Disponibilidade de artefatos

Os artefatos deste estudo estão disponíveis no link: <https://anonymous.4open.science/r/sti-calculo-sbie2026-0E2E/>.

## Referências

- [Baker and Inventado 2014] Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. In Larusson, J. A. and White, B., editors, *Learning Analytics: From Research to Practice*, pages 61–75. Springer, New York.
- [BRASIL 2023] BRASIL (2023). Notas sobre o brasil no pisa 2022.
- [Carbonell 1970] Carbonell, J. R. (1970). Ai in cai: An artificial-intelligence approach to computer-assisted instruction. In *IEEE Transactions on Man-Machine Systems*, volume 11, pages 190–202.
- [Cheng et al. 2025] Cheng, C., Huang, Z., Zhao, G., Guo, Y., Lin, X., Wu, J., Li, X., and Wang, S. (2025). From objectives to questions: A planning-based framework for educational mathematical question generation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12836–12856, Vienna, Austria. Association for Computational Linguistics.
- [Feng et al. 2024a] Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., Ornelas, N., and Lan, A. (2024a). Exploring automated distractor generation for math multiple-choice questions via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3067–3082, Mexico City, Mexico. Association for Computational Linguistics.
- [Feng et al. 2024b] Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., Ornelas, N. O., and Lan, A. (2024b). Exploring automated distractor generation for math multiple-choice questions via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.
- [Ferreira-Mello et al. 2019] Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

- [Grévisse et al. 2024] Grévisse, C., Pavlou, M. A. S., and Schneider, J.-G. (2024). Docimological quality analysis of llm-generated multiple choice questions in computer science and medicine. *SN Computer Science*.
- [Kasneci et al. 2023a] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023a). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- [Kasneci et al. 2023b] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023b). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- [Kurdi et al. 2020a] Kurdi, G., Leo, J., Parsia, B., et al. (2020a). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.
- [Kurdi et al. 2020b] Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020b). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.
- [Kurdi et al. 2020c] Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020c). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.
- [Lee et al. 2024a] Lee, J., Smith, D., Woodhead, S., and Lan, A. (2024a). Math multiple choice question generation via human-large language model collaboration. In *Proceedings of the 17th International Conference on Educational Data Mining*.
- [Lee et al. 2024b] Lee, J., Smith, D., Woodhead, S., and Lan, A. (2024b). Math multiple choice question generation via human-large language model collaboration. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 941–946, Atlanta, Georgia, USA. International Educational Data Mining Society.
- [Nogueira et al. 2025] Nogueira, J. V. d. C. M. F., Castro, J. A. P. d., Larcher, L. O., Ferreira, R. V. J., Barbosa, B. T., and Souza, J. F. d. (2025). Uma análise de qualidade do uso de grandes modelos de linguagem para geração automática de itens avaliativos em português. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 235–248. SBC.
- [Quincozes et al. 2025] Quincozes, C. B., Molinos, D., Araújo, R. D., Quincozes, S., and Guedes, G. T. A. (2025). Engenharia de prompt para a geração automatizada de questões assistida por llms: Uma análise comparativa. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 1347–1360. SBC.
- [Santos et al. 2025] Santos, M., Barros, A., Santos, E., Silva, J., Isotani, S., Bittencourt, I., Macario, V., Rodrigues, L., and Dermeval, D. (2025). Geração de questões com llms leves: Um estudo inicial sobre a percepção de educadores. In *Anais do XXXVI*

*Simpósio Brasileiro de Informática na Educação*, pages 1636–1647, Porto Alegre, RS, Brasil. SBC.

- [SANTOS et al. 2024] SANTOS, R. D. d., MAGALHÃES, W. B. C., ANJOS, A. L. C. d., and LETTI, A. G. (2024). Evasão e retenção dos cursos de licenciatura em matemática na bahia no período de 2010 a 2022. In *Anais do Seminário Nacional de Políticas Públicas, Gestão e Práxis Educacional*, pages 189–199, Vitória da Conquista.
- [Scarlato et al. 2024a] Scarlato, A., Feng, W., Smith, D., Woodhead, S., and Lan, A. (2024a). Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank. In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z., editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 222–231, Mexico City, Mexico. Association for Computational Linguistics.
- [Scarlato et al. 2024b] Scarlato, A., Feng, W., Smith, D., Woodhead, S., and Lan, A. (2024b). Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 222–231, Mexico City, Mexico. Association for Computational Linguistics.
- [Tall 1993] Tall, D. (1993). Students’ difficulties in calculus. In *Proceedings of Working Group 3 on Students’ Difficulties in Calculus*, pages 13–28.
- [VanLehn 2011] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.
- [Wang et al. 2025] Wang, L., Song, R., Guo, W., and Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3):2559–2584.
- [Woelfel 2009] Woelfel, B. P. (2009). *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning*. Morgan Kaufmann, Burlington.
- [Yan et al. 2024] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.