VII SIMAC
SIMPÓSIO ACADÊMICO
FACULDADE ENG. SALVADOR ARENA

Confiabilidade dos dados em inteligência artificial generativa na educação: desafios, percepções e soluções emergentes

Daniele Soares Cardoso, 071240043@faculdade.cefsa.edu.br, Faculdade Engenheiro Salvador Arena
Victor Flohr Costa Bicudo Larrubia, 082210026@faculdade.cefsa.edu.br, Faculdade Engenheiro Salvador Arena
Victor Inácio Oliveira, pro14724@cefsa.edu.br, Faculdade Engenheiro Salvador Arena

Resumo

A difusão de sistemas de geração automática de texto em contextos educacionais amplia oportunidades de acesso e personalização, mas impõe a necessidade de garantir fidedignidade e rastreabilidade das respostas. Este estudo, em andamento, investiga como métricas de avaliação e arranjos de verificação podem reduzir alucinações e qualificar a auditabilidade de modelos de linguagem no uso pedagógico. O método integra: análise crítica de métricas voltadas a forma e semântica (ROUGE, BERTScore, BLEURT) e a veracidade factual (QAGS, FactScore, HHEM); mapeamento de tarefas educacionais típicas (sínteses explicativas, respostas conceituais e geração de questões); e um *pipeline* com RAG (recuperação e citação de evidências), XAI (explicações sobre o processo decisório) e validação multiagente (checagens paralelas e consenso). Os resultados preliminares indicam que nenhuma métrica isolada cobre, simultaneamente, fluência, coerência e veracidade; combinações que articulam avaliadores de factualidade com indicadores semânticolinguísticos, ancoradas em fontes e acompanhadas de justificativas, produzem diagnósticos mais estáveis e úteis para a prática docente. Propõe-se, por fim, um protocolo de adoção que prioriza triagem de alucinação (HHEM), confirmações adicionais (QAGS/FactScore) e monitoramento de clareza (BERTScore/BLEURT), sempre com suporte de RAG e XAI.

Palavras-chave: Inteligência artificial generativa. Confiabilidade. Alucinações. Educação. Métricas de avaliação.

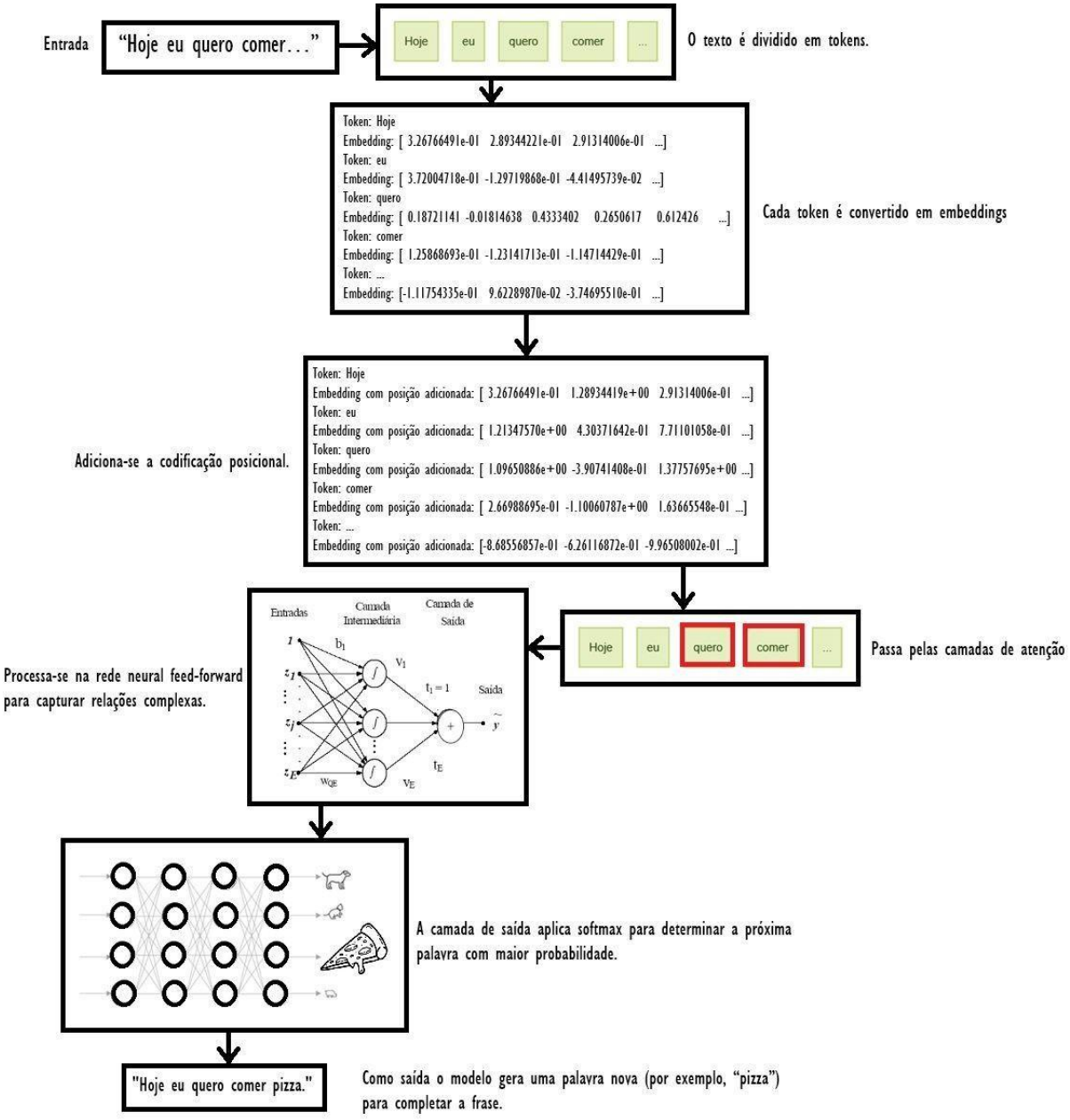
Introdução

Modelos generativos passaram a compor rotinas de estudo, autoria e revisão de textos em ambientes formais de ensino (BROWN et al., 2020; POPENICI; KERR, 2017; SHOUFAN, 2023). A utilidade desses sistemas vem acompanhada de um desafio central: garantir qualidade informacional quando as respostas não trazem, por padrão, as bases empíricas que as sustentam. Em atividades avaliativas, produção de materiais didáticos e apoio à pesquisa, erros factuais, vieses e falta de justificativas podem comprometer a integridade formativa (ROYAL SOCIETY, 2023; MOTOKI; ALMEIDA; VARGAS, 2023; CARVALHO, 2021).

Este artigo sintetiza os elementos essenciais de uma investigação sobre confiabilidade dos dados em modelos generativos aplicados à educação. O objetivo é oferecer um enquadramento técnico-metodológico

que permita avaliar e mitigar riscos, combinando métricas complementares e mecanismos de validação. As contribuições concentram-se em três frentes: (i) delimitação do papel de cada métrica e de seus limites (WANG; CHO; LEWIS, 2020; MIN et al., 2023; GHOSH, 2024); (ii) proposta de uma arquitetura que integra RAG e XAI com validação multiagente (ROYAL SOCIETY, 2023; DORRI; KANHERE; JURDAK, 2018; PREMNATH; YAU, 2023); e (iii) um protocolo operacional para uso pedagógico responsável, atento a custos, cobertura linguística e governança (UNESCO, 2021).

Figura 1 – ilustrando as camadas de uma IAG Trasformer



Fonte: Autoria própria (2025)

Metodologia

O percurso metodológico articula três etapas. A primeira consiste na caracterização das métricas mais empregadas para avaliar saídas de modelos de linguagem: ROUGE (sobreposição lexical) (LIN, 2004); BERTScore (similaridade semântica) (ZHANG et al., 2019); BLEURT (qualidade linguística com sinal humano) (SELLAM; DAS; PARIKH, 2020); QAGS (cheque de consistência factual via QG/QA) (WANG; CHO; LEWIS, 2020); RQUGE (avaliação de perguntas geradas) (MOHAMMADSHAHI et al., 2023); FactScore (verificação de fatos por extração e inferência) (MIN et al., 2023); e HHEM (detecção supervisionada de alucinação) (GHOSH, 2024). Em seguida, define-se um conjunto de tarefas educacionais representativas — síntese de conteúdos a partir de materiais de aula, resposta a questões conceituais e geração de itens — para observar adequação e limitações das métricas em cenários reais. Por fim, desenha-se um pipeline integrado com RAG, XAI e comitê multiagente, de modo a ancorar respostas em evidências, explicitar o raciocínio do sistema e distribuir verificações entre agentes especializados (factualidade, clareza e conformidade pedagógica) (ROYAL SOCIETY, 2023; DORRI; KANHERE; JURDAK, 2018; KAMBLE et al., 2022; PREMNATH; YAU, 2023).

A estratégia privilegia triangulação. Avaliadores de forma/semântica (ROUGE, BERTScore, BLEURT) são combinados a medidores de veracidade (QAGS, FactScore, HHEM). RAG fornece lastro documental com citações de trechos, enquanto XAI explicita critérios e passos decisórios. Em conjunto, essas camadas aproximam a avaliação do que interessa à prática pedagógica: conteúdo correto, compreensível e verificável (ROYAL SOCIETY, 2023; UNESCO, 2021).

Resultados preliminares

A comparação entre métricas revela funções complementares. ROUGE oferece um indicador rápido de alinhamento ao conteúdo esperado, mas não captura sentido nem consistência factual. BERTScore é sensível a paráfrases e aproximações semânticas, enquanto BLEURT evidencia fluência e naturalidade; ambos, porém, não atestam veracidade. Para fatos, QAGS e FactScore mostram maior aderência, embora exijam infraestrutura e impliquem maior custo computacional. HHEM destaca-se como triagem eficaz de alucinações quando há contexto explícito de verificação, funcionando como barreira inicial antes de inspeções mais dispendiosas.

À luz desses achados, propõe-se um protocolo de uso para situações de sala de aula. Em atividades em que a veracidade é crucial — explicações conceituais, sínteses com apoio em materiais e orientações de estudo —, aplicar HHEM como filtro inicial de alucinações; quando necessário, confirmar com QAGS ou FactScore para exame granular de trechos. Em paralelo, monitorar clareza expositiva e coerência semântica por BERTScore e BLEURT. Em tarefas orientadas a gabaritos ou resumos de alto nível, ROUGE permanece útil como apoio, desde que não seja critério único.

A integração de RAG e XAI agrega duas virtudes pedagógicas. A ancoragem em fontes reduz inventivas e facilita auditoria por docentes e estudantes; a explicabilidade aproxima a prática do método científico, ao tornar transparentes as razões pelas quais certos trechos foram selecionados e como a conclusão foi alcançada. A validação multiagente eleva a robustez: agentes com papéis distintos executam checagens em paralelo e deliberam por consenso, mitigando a dependência de um avaliador único. Entre as limitações, destacam-se custo computacional e latência quando verificações profundas são acionadas em larga escala; esses efeitos podem ser administrados com amostragem, limiares para disparo de verificações

pesadas e cache de evidências em tarefas recorrentes. Persiste, ainda, o desafio da cobertura linguística, dado que parte dos avaliadores foi otimizada para o inglês, exigindo adaptações ao português e a domínios específicos.

Tabela 1 – Métricas em aplicações relacionadas a área da educação

Situação Educacional	Métrica(s) Recomendada(s)	Justificativa
Avaliação da precisão factual de uma explicação de IA com base em material didático	FactScore, QAGS, HHEM	Avaliam se as informações estão corretamente ancoradas em textos de referência
Verificação de erros conceituais em resumos gerados por IA	HHEM, QAGS	Detectam alucinações factuais e inconsistências por meio de inferência e QA
Checagem de similaridade entre textos gerados e gabaritos humanos	BERTScore, ROUGE	Úteis como métricas complementares em tarefas de correspondência geral
Construção de um sistema automatizado de verificação em tempo real	HHEM + FactScore integrados	Permite sinalizar respostas potencialmente alucinadas durante a interação
Monitoramento contínuo da qualidade de respostas de IA em plataformas educacionais	QAGS + BLEURT	Combina avaliação factual com análise de clareza e estrutura textual

Fonte: Autoria própria (2025)

Resultados preliminares

Este estudo sintetiza um arranjo operacional para elevar a confiabilidade de sistemas generativos em contextos educacionais, articulando métricas complementares (forma/semântica e factualidade) com RAG (ancoragem em fontes), XAI (explicações legíveis) e validação multiagente (checagens paralelas e consenso). O conjunto favorece auditabilidade, transparência e adequação pedagógica, reduzindo a incidência de alucinações e fornecendo critérios explícitos para docentes e equipes técnicas.

Como pesquisa em andamento, as conclusões são provisórias. Permanecem desafios de custo e latência quando verificações profundas são acionadas, além de cobertura linguística desigual — pontos que demandam otimizações (amostragem, limiares, cache de evidências) e adaptação ao português e a domínios específicos. Em termos institucionais, a efetividade do protocolo depende de governança clara (papéis, parâmetros e responsabilidades) e de formação docente para leitura crítica das saídas de IA.

Os próximos passos incluem: estabelecer limiares operacionais por tarefa (faixas de HHEM e QAGS para resumos, QA e geração de questões), ampliar a avaliação em cenários reais com docentes e estudantes, e mensurar trade-offs entre precisão, tempo de resposta e recursos computacionais. Ao consolidar esses elementos, o roteiro proposto tende a viabilizar um uso mais seguro, verificável e pedagogicamente útil das IAGs na educação.

Referências

ALMEIDA, J. P.; PRADO, R. V. **Aplicações éticas da inteligência artificial na ciência aberta: confiabilidade e integridade dos dados**. Revista Brasileira de Ciência da Informação, v. 15, n. 2, p. 103-119, 2023.

BROWN, Tom B. et al. **Language Models are Few-Shot Learners**. arXiv, 2020. Disponível em: https://arxiv.org/abs/2005.14165. Acesso em: 23 jun. 2025.

CARVALHO, André Carlos Ponce de Leon Ferreira de. **Inteligência artificial: riscos, benefícios e uso responsável**. **Estudos Avançados**, São Paulo, v. 35, n. 101, p. 21–35, 2021. DOI: 10.1590/s0103-4014.2021.35101.003. Disponível em: https://revistas.usp.br/eav/article/view/185020/171203. Acesso em: 9 abr. 2025.

DORRI, A.; KANHERE, S. S.; JURDAK, R. **Multi-Agent Systems: A Survey**. IEEE Access, v. 6, p. 28573–28593, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2831222.

GHOSH, Shalini. **Hallucination Evaluation Model (HHEM)**. Vectara / Hugging Face, 2024. Disponível em: https://huggingface.co/vectara/hallucination_evaluation_model. Acesso em: 23 jun. 2025.

KAMBLE, S. et al. **Swarm Intelligence for Fake News Detection**. Nature Machine Intelligence, v. 4, n. 2, p. 136–145, 2022. DOI: https://doi.org/10.1038/s42256-021-00437-7.

KRYŚCIŃSKI, Wojciech et al. **Evaluating the Factual Consistency of Abstractive Text Summarization**. arXiv, [s.d.]. Disponível em: https://arxiv.org/abs/2004.04228. Acesso em: 23 jun. 2025.

LIN, Chin-Yew. **ROUGE: A Package for Automatic Evaluation of Summaries**. In: *Text Summarization Branches Out*. Barcelona: ACL, 2004. p. 74–81. Disponível em: https://aclanthology.org/W04-1013.pdf. Acesso em: 21 ago. 2025.

MIN, Sewon et al. **FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long-Form Text Generation**. In: *EMNLP 2023*. 2023. Disponível em: https://aclanthology.org/2023.emnlp-main.741/. Acesso em: 21 ago. 2025.

MOTOKI, Kauê; ALMEIDA, Luana; VARGAS, Márcia. **Geração de conteúdo e viés: uma análise crítica das respostas de IAs generativas em língua portuguesa**. Revista Brasileira de Tecnologias Educacionais, v. 4, n. 2, p. 45–60, 2023.

MOHAMMADSHAHI, Alireza et al. **RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question**. *Findings of ACL 2023*, 2023. DOI: 10.18653/v1/2023.findings-acl.428. Disponível em: https://aclanthology.org/2023.findings-acl.428.pdf. Acesso em: 21 ago. 2025.

PATEL, Keyur; KELLY, Maria; FERNANDEZ, Jorge. **Generative AI in Education: Promise and Perils**. International Journal of Educational Technology, v. 9, n. 1, p. 10–24, 2023.

POPENICI, Stefan A. D.; KERR, Sharon. **Exploring the impact of artificial intelligence on teaching and learning in higher education**. Research and Practice in Technology Enhanced Learning, v. 12, n. 1, p. 1–13, 2017. DOI: https://doi.org/10.1186/s41039-017-0062-8.

PREMNATH, Shashwat; YAU, K.-L. A. **Decentralized AI Verification Using Swarm Learning**. ACM Computing Surveys, v. 55, n. 4, p. 1–35, 2023. DOI: https://doi.org/10.1145/3558483.

ROYAL SOCIETY. **Explainable AI: The new frontier**. Royal Society Open Science, 2023. Disponível em: https://royalsocietypublishing.org/doi/epdf/10.1098/rsos.230658. Acesso em: 9 abr. 2025.

SELLAM, Thibault; DAS, Dipanjan; PARIKH, Ankur. **BLEURT: Learning Robust Metrics for Text Generation**. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. p. 7881–7892, 2020. DOI: 10.18653/v1/2020.acl-main.704. Disponível em: https://aclanthology.org/2020.acl-main.704/. Acesso em: 21 ago. 2025.

SHOUFAN, Adel. **The Reliability of AI in Education: Between Innovation and Illusion**. Computers & Education: Artificial Intelligence, v. 4, 2023. DOI: https://doi.org/10.1016/j.caeai.2023.100094..

SILVA, M. A.; FONSECA, R. L.; GOMES, C. P. **Inteligência artificial e validação científica: desafios para a confiabilidade dos dados**. Revista Brasileira de Informação em Ciência e Tecnologia, v. 11, n. 1, p. 78-95, 2022.

UNESCO. **Recommendation on the Ethics of Artificial Intelligence**. Paris: UNESCO, 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137. Acesso em: 9 abr. 2025.

VASWANI, Ashish et al. **Attention is All You Need**. NeurIPS, 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 23 jun. 2025.

WANG, Alex; CHO, Kyunghyun; LEWIS, Mike. **Asking and Answering Questions to Evaluate the Factual Consistency of Summaries (QAGS)**. In: *Proceedings of ACL 2020*. 2020. Disponível em: https://aclanthology.org/2020.acl-main.450/. Acesso em: 21 ago. 2025.

ZHANG, Tianyi et al. **BERTScore: Evaluating Text Generation with BERT**. In: *ICLR 2020*. 2019 (preprint). Disponível em: https://arxiv.org/abs/1904.09675. Acesso em: 21 ago. 2025.