O desempenho de grandes modelos de linguagem na síntese de relatos de pesquisa educacional: Estudo comparativo com artigos no idioma português

Ronei Ximenes Martins, Maria Kamylla Silva Xavier, Bruno Amarante Couto Rezende, Nicole de Santana Gomes Garcia, Patricia Peixoto Carneiro Viegas, Francine de Paulo Martins Lim

https://doi.org/10.1590/SciELOPreprints.17611

Submetido em: 2026-08-23
Postado em: 2026-09-28 (versão 2)
(AAAA-MM-DD)

Justificativa da versão: Esta versão foi atualizada em 26/9/2026 para correção de erros na digitação das estatísticas dos testes de Kruskal-Wallis destinados à comparação de resultados por modelo de IAG e tipo de prompt. Também foi inserida, no Quadro 1, a íntegra do conteúdo dos prompts utilizados e foi ajustado o tópico dos resultados referentes à análise baseada nos tipos de prompt, pois os gráficos das figuras 2 e 3 continham imprecisões. Eles foram substituídos por uma tabela. Os ajustes não invalidam a versão anterior no que se refere às conclusões ou as discussões decorrentes da interpretação dos resultados.

O desempenho de grandes modelos de linguagem na síntese de relatos de pesquisa educacional: Estudo comparativo com artigos no idioma português[1]

Ronei Ximenes Martins
Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil. ORCID: https://orcid.org/0000-0002-3586-9086

Maria Kamylla e Silva Xavier
Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brasil. ORCID: https://orcid.org/0000-0001-8602-1870

Bruno Amarante Couto Rezende
Centro Federal de Educação Tecnológica de Minas Gerais, Varginha, Minas Gerais, Brasil. ORCID: https://orcid.org/0000-0002-5076-8080

Nicole de Santana Gomes Garcia
Universidade do Estado de Minas Gerais, Campanha, Minas Gerais, Brasil. ORCID: https://orcid.org/0000-0002-9107-7016

Patrícia Peixoto Carneiro Viegas
Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil. ORCID: https://orcid.org/0000-0002-1341-0098

Francine de Paulo Martins Lima
Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil. ORCID: https://orcid.org/0000-0002-9646-8235

RESUMO: Este estudo objetiva comparar o desempenho do ChatGPT 5.0, Gemini 3.0 Flash e DeepSeek V3 na síntese de artigos científicos publicados em português. Adotou-se abordagem mista, com teste de geração de resumos para cinco artigos da área da educação em três variações de prompt. A análise dos *outputs* foi realizada por 15 especialistas nas temáticas dos artigos que utilizaram uma escala ordinal de Percepção de Qualidade (PQ) para mensurar as dimensões de qualidade de conteúdo e isenção de vieses. Os resultados indicaram bom desempenho dos modelos, com resultados de PQ entre 84% e 94% do escore total. O teste de *Kruskal-Wallis* identificou diferenças significativas na comparação dos resultados por artigo, ($H(4) = 24{,}040$, $p < 0{,}001$) e inexistência de diferenças entre os modelos ($H(2) = 1{,}604$, $p = 0{,}448$) ou tipos de prompt ($H(8) = 2{,}506$, $p = 0{,}961$), sugerindo paridade tecnológica na tarefa de síntese de textos científicos. Ao lado disso, a análise qualitativa identificou tendência à generalização e à supressão de marcadores de incerteza, evidências que indicam falhas na validade dos resumos produzidos. Os resultados obtidos também sugerem potenciais vieses oriundos do idioma de treinamento dos modelos, que, hipoteticamente, podem afetar a captura da textura narrativa e de jargão específico da ciência

[1] Esta versão foi atualizada em 26/9/2026 para correção de erros na digitação das estatísticas dos testes de *Kruskal-Wallis* destinados à comparação de resultados por modelo de IAG e tipo de prompt. Também foi inserida, no Quadro 1, a íntegra do conteúdo dos prompts utilizados e foi ajustado o tópico dos resultados referentes à análise baseada nos tipos de prompt, pois os gráficos das figuras 2 e 3 continham imprecisões. Eles foram substituídos por uma tabela. Os ajustes não invalidam as conclusões ou as discussões decorrentes da interpretação dos resultados da versão anterior.

produzida em português. Concluiu-se que, embora os modelos de IAG testados apresentem bom desempenho sintático, a supervisão humana permanece indispensável para garantia da fidedignidade e integralidade da comunicação científica da pesquisa em educação.

Palavras-chave: Inteligência Artificial Generativa, Comunicação científica, Processamento de resumos científicos, Vieses Algorítmicos.

The Performance of Large Language Models in Synthesising Educational Research Reports: A Comparative Study of Articles in Portuguese

ABSTRACT: This study compares the performance of ChatGPT 5.0, Gemini 3.0 Flash, and DeepSeek V3 in producing abstracts of scientific articles published in Portuguese. A mixed-methods approach was adopted, involving the submission of five articles in the field of Education to the three language models using three prompt variations. The outputs were evaluated by 15 specialists, with an ordinal perception of quality (PQ) scale to assess the dimensions of content quality and freedom from bias. The results indicated that the models performed well, with PQ medians ranging from 84% to 94% of the total score. The Kruskal–Wallis test identified statistically significant differences when the results were compared by article, $H(4)=24.040, p<0.001$, but found no significant differences among the models, $H(2)=1.604, p=0.448$, or among the prompt types, $H(8) = 2,506$, $p = 0,961$. These findings suggest a technological convergence among the models in the task of scientific text synthesis. However, the qualitative analysis identified generalizations' tendency and the suppression of uncertainty markers, indicating shortcomings in the validity of the generated abstracts. The findings also suggest the possibility of biases arising from the languages represented in the models' training data, which could hypothetically affect their ability to capture the narrative texture and specialized terminology of scientific knowledge produced in Portuguese. It was concluded that, although the generative AI models tested demonstrated good syntactic performance, human oversight remains indispensable for ensuring the accuracy and completeness of scientific communication in educational research.

Keywords: Generative Artificial Intelligence, Scientific Communication, Scientific Abstract Processing, Algorithmic Biases.

1. INTRODUÇÃO

A utilização de sistemas de Inteligência Artificial (IAG) baseados em *Large Language Models* (LLMs) no contexto acadêmico é foco de debates, com manifestações que vão da rejeição radical ao incentivo para utilização em etapas específicas do ensino, da pesquisa e da extensão. Dados recentes sobre o contexto educacional brasileiro revelam que 70% dos estudantes do Ensino Médio já utilizam ferramentas de LLM em suas pesquisas escolares, embora apenas 32% confirmam ter recebido orientação formal sobre esse uso. Entre os professores, a adesão também é expressiva, com 43% dos professores declarando o uso dessas tecnologias para a preparação de conteúdos didáticos (CGI.BR, 2025).

Essa disseminação, iniciada no Brasil a partir do segundo semestre de 2022, suscita incertezas sobre as consequências do uso no campo educacional e da pesquisa científica. No plano internacional, enquanto algumas universidades publicaram diretrizes para uso ético e transparente da IA generativa, outras adotaram proibições amplas em avaliações ou em

disciplinas específicas, com enquadramento disciplinar rigoroso por violação de integridade acadêmica (Crawford; Cowling; Allen, 2023; Xiao; Chen; Bao, 2023)

Os que incentivam o uso se apoiam no potencial da IAG em atuar como um assistente que amplia a produtividade na escrita, análise de dados e revisão de literatura (Sampaio et al., 2024). Em contrapartida, as manifestações de rejeição se baseiam em riscos tais como a falta de confiabilidade das informações (alucinações), a presença de vieses algorítmicos, o perigo de plágio e a ameaça aos princípios de integridade acadêmica (Lima; Felippe, 2025). Diante desse cenário de potencial polarização, Lima e Felipe (op cit.) defendem a urgência de uma mediação pedagógica e estruturada, uma vez que a adoção individual tem avançado de forma muito mais rápida do que a implementação de salvaguardas e práticas responsáveis pelas instituições.

Especificamente no que se refere à aplicação em pesquisas do campo educacional, há, na literatura, lacunas de compreensão sobre os efeitos na forma como o conhecimento acadêmico é processado e disseminado. Ferramentas como *ChatGPT, Gemini* e *DeepSeek* têm sido adotadas por pesquisadores, estudantes e instâncias de comunicação científica para analisar relatos de pesquisa, apoiar revisão bibliográfica e estruturar sínteses de conjuntos de artigos científicos. Contudo, a delegação dessa tarefa a sistemas automatizados levanta preocupações críticas quanto à precisão, fidedignidade factual e responsabilidade ética na transposição dos dados originais para os resumos e sínteses gerados.

Estudos recentes, como os de Peters e Chin-Yee (2025) e Lima e Felipe (2025), apontam limitações na produção de textos científicos por essas ferramentas. O estudo de Peters e Chin-Yee, especificamente, demonstra que LLM populares em uso apresentam viés sistêmico de generalização ao resumirem textos científicos em inglês, produzindo ampliação de resultados presentes nos textos originais em aproximadamente 70% dos casos testados. Tal fenômeno se manifestou, segundo Peters e Chin-Yee, pela transposição de resultados limitados observados em tempo passado para aplicáveis ao presente, bem como a transformação indevida de descrições de resultados em recomendações práticas orientadas à ação. Estas distorções são preocupantes, pois a generalização pode induzir a interpretações incorretas de resultados e seus efeitos, com a consequente criação de expectativas indevidas e evidências distorcidas.

No contexto brasileiro, onde o acesso à informação científica em língua portuguesa é crucial para a democratização do conhecimento, se torna relevante investigar se os padrões de viés identificados na literatura internacional estão presentes de forma similar em nossa língua e cultura científica. Embora a existência de vieses de generalização em LLMs para textos em inglês já esteja documentada, há uma lacuna de conhecimento sobre a existência e as características desses vieses na produção automática de resumos de artigos científicos lusófonos. Neste trabalho, hipotetiza-se que diferenças linguísticas e culturais, somadas às especificidades da produção científica brasileira no campo educacional, bem como as diferenças no treinamento dos LLM com fontes em inglês e em português, podem gerar interações não desejadas com os algoritmos desses modelos, afetando diretamente a qualidade do output.

Ante o exposto, esta pesquisa teve como objetivo comparar o desempenho de diferentes IAG (*ChatGPT, Gemini e DeepSeek*) para a produção de resumos de relatórios de pesquisa publicados em português brasileiro. A investigação buscou avaliar como esses modelos se comportam na tarefa específica de comunicar com precisão o problema, o objetivo, a metodologia, os resultados e as conclusões de artigos científicos, identificando possíveis diferenças em suas capacidades de síntese e manutenção fiel das informações relatadas nos textos originais.

2. REFERENCIAL TEÓRICO

O campo do Processamento de Linguagem Natural (PLN) no âmbito da IAG e a subjacente aplicação à comunicação científica tem relação direta com a transição dos modelos de aprendizado de máquina baseados no paradigma discriminativo para os ancorados no paradigma generativo. Esta mudança de paradigma afetou a forma e o conteúdo produzidos, incluindo a comunicação científica.

Conforme descrito por Jurafsky e Martin (2026), no paradigma discriminativo, característico de classificadores como a regressão logística, o foco reside no cálculo da probabilidade condicional para categorizar dados a partir de recursos extraídos do input. Isso os torna eficientes em ações de classificação, mas apresenta limitação na produção de novos conteúdos. Em contrapartida, o paradigma generativo, como os modelos de n-gramas[2], atribui probabilidades a sequências de palavras. Atualmente, este paradigma é consolidado pelos Grandes Modelos de Linguagem (LLM), que utilizam a predição probabilística da próxima palavra para permitir a síntese e a geração condicional de novos conteúdos linguísticos. O aumento recente no uso de modelos generativos está relacionado ao crescimento da escala de dados de treinamento e parâmetros, que o tornou operacionalmente viável para a produção textual.

O trabalho de Bender et al (2021) apresentou um panorama evolutivo desta tecnologia até se chegar aos modelos com arquitetura *Transformer*[3], proposta por Vaswani et al. (2017) e que se beneficiaram do aumento da escala e volume de dados de treinamento. Para Bender et al (2021), os modelos *Transformer* possuem capacidade de executar, a partir de poucos exemplos, tarefas aparentemente sensíveis ao significado, entre elas a síntese de textos e a resposta a perguntas. Neste sentido, Brown et al. (2020) argumentam que essa abordagem de execução pode necessitar de ajustes específicos, seja com poucos exemplos (*few-shot*) ou, mesmo sem exemplo (*zero-shot*).

Portanto, as mudanças trazidas pelos modelos *Transformers* estão relacionadas à capacidade de conseguir prever o próximo *token* em textos longos e com síntese textual aparentemente fluida e coerente. Isto tornou os LLM potencialmente aplicáveis também para a produção da comunicação científica. No entanto, Kaufman (2022) adverte que essa capacidade não deve ser interpretada como criatividade ou compreensão por parte das máquinas, mas sim como um salto computacional e estatístico, cuja habilidade de gerar novos conteúdos se origina da modelagem matemática de padrões linguísticos e não da compreensão semântica refinada.

Ao lado disso, é preciso destacar que esses modelos de linguagem não possuem qualquer forma de discernimento ou competência epistêmica. Bender e colaboradores (2021) classificam os LLMs como "papagaios estocásticos" (*stochastic parrots*) com a justificativa de que eles criam informação a partir de cálculos probabilísticos sequências sem ancoragem em significados reais, compromisso com a verdade ou intencionalidade semântica.

Por outro lado, estudos como os de Brown et al. (2020) e Wei et al. (2022) trabalham com a hipótese de que o aumento expressivo nos parâmetros e no volume de dados de treinamento permitem que LLMs adquira e demonstre conhecimento semântico e de mundo. Brown e colaboradores (2020), ao discutirem sobre o ChatGPT 3, afirmam que esse aumento

[2] Modelo probabilístico mais antigo que estima a probabilidade de uma palavra ou token com base nas n−1 palavras anteriores (Bender et al., 2021 e Jurafsky e Martin, 2026).

[3] Arquitetura padrão utilizada na construção de LLM. Trabalha com camada de *self-attention*, um mecanismo que constrói representações contextualizadas do significado de cada token ao ponderar e integrar informações dos outros tokens (Jurafsky e Martin, 2026).

nas escalas permitiu que os modelos resolverem questões semânticas complexas e aplicassem raciocínio de senso comum para gerar outputs. Wei et al. (2022) consideram essa capacidade uma "habilidade emergente", apoiados na premissa de que a capacidade de interpretar contextos é uma mudança qualitativa que emerge quando os modelos atingem escalas muito grandes de parâmetros.

Entretanto, quando aplicados ao campo da comunicação científica, os resultados (outputs) dos LLM merecem mais atenção, pois ela é fundamental no processo de disseminação do novo conhecimento produzido pelas pesquisas e requer exatidão de significado. A comunicação científica fundamenta os novos processos de pesquisa e o ensino formal. Neste contexto, a síntese científica, definida aqui como o processo de condensar e resumir textos científicos preservando suas características e resultados, é prática corriqueira, principalmente no ensino superior e na escrita acadêmica. Até a popularização do uso da IAG para esta prática, era uma atividade realizada eminentemente por humanos, mas desde 2022 tem se intensificado a utilização de modelos baseados em LLM no apoio a diferentes etapas da comunicação científica.

Retomando a metáfora dos "papagaios estocásticos", é necessário destacar que os modelos de LLMs não compreendem os significados da linguagem presente no artigo a ser resumido. Bender et al. (2021) citam que humanos têm uma predisposição para interpretar textos, mesmo os gerados por máquinas, como se tivessem sentido e intenção comunicativa, assim o leitor pode atribuir profundidade intelectual e rigor conceitual a um resultado que é, em essência, um produto estatístico.

Ferramentas como o ChatGPT, o Gemini e o DeepSeek processam rapidamente grandes quantidades de informações complexas, com baixo custo para o usuário e, aparentemente, oferecem resumos plausíveis de relatos de pesquisa. Esse processo tem potencial para auxiliar pesquisadores, estudantes e professores mas existe um risco associado. Mesmo criando sínteses textuais fluidas e textualmente coerentes, os outputs podem omitir ou valorizar partes de resultados, incertezas observadas, limitações ou conclusões de uma pesquisa (Peters e Chin-Yee, 2025).

O processo de síntese com suporte automatizado, por si, também passa por uma mudança de paradigma, segundo Peters e Chin-Yee (2025). Se antes era baseado em procedimentos extrativos, com recortes de frases literais do texto original, os modelos de LLM introduziram um novo formato, abstrato, com processamento que utiliza o texto completo para recriar uma descrição textual resumida, aparentemente fluida e textualmente coerente, do que foi comunicado no relato completo. Contudo, a questão central é até que ponto se pode confiar no que, aparentemente, é elegante e coerente.

O estudo desenvolvido por Peters e Chin-Yee (2025) analisou 4.900 resumos em língua inglesa gerados por 10 modelos de LLMs. Os resultados indicaram vieses nos outputs da maioria destes LLMs, nos quais ocorreu o que os autores chamaram de supergeneralização em relação às conclusões originais dos estudos científicos. O experimento identificou que modelos mais novos, como o ChatGPT 4 e o LLaMA 3.3, introduziram generalizações algorítmicas em até 73% dos resumos gerados.

A questão da generalização citada por Peters e Chin-Yee (2025) pode ser ainda mais complexa se faz um recorte da síntese de comunicação científica em língua portuguesa. O treinamento dos modelos de LLMs mais utilizados é realizado com dados não revisados e retirados da internet, a maioria em língua Inglesa, tornando esta a língua padrão dos referidos modelos. Existem iniciativas de LLMs em Língua Portuguesa, como o Sabiá (Pires; Abonizio; Nogueira, 2023), ainda limitados em alcance e adoção se comparado aos modelos internacionais.

Ante o exposto, para compreender possíveis dos modelos de LLM na síntese de relatos de pesquisa em português, nesta pesquisa selecionou-se três IAG amplamente adotadas, cuja natureza interna de arquitetura e treinamento diferem em função das empresas que as criaram: o *Chat GPT (OpenAI)*, o *Gemini (Google)* e o *DeepSeek* (da chinesa *DeepSeek-AI*). A seleção desses modelos de linguagem para o estudo comparativo não foi arbitrária, o critério adotado foi o de predominância de utilização no Brasil, no caso das duas primeiras (fundação itaú; Datafolha, 2025) e a DeepSeek como contraponto de filosofia de construção, considerando que esta diferença tem potencial para impactar a qualidade e a fidedignidade dos *outputs* em português.

Os principais modelos de linguagem disponíveis na atualidade compartilham um processo de desenvolvimento semelhante. Todos passam por um treinamento inicial com grandes volumes de texto, seguido por duas etapas de ajuste que moldam seu comportamento: (i) o Ajuste Fino Supervisionado (SFT), em que o modelo aprende com exemplos humanos de como responder; e (ii) o Aprendizado por Reforço com *Feedback* Humano (RLHF), por meio do qual as respostas são avaliadas e refinadas para se tornarem mais úteis e seguras (Brown *et al*., 2020; Ouyang *et al*., 2022). O problema desse tipo refinamento é que orienta o modelo para dar respostas diretas e assertivas. O resultado são textos fluentes e convincentes, mas com risco elevado de simplificação ou generalizações indevidas.

Especificamente em relação às IAG testadas. Além os ajustes indicados no parágrafo anterior, observa-se que, no caso do Chat GPT, segundo a OPENAI (2023), este apresenta ajustes centrados em comportamento com tendência a priorizar respostas rápidas e conclusivas. Destaca-se que este é um ponto crítico para a produção de textos científicos, nos quais a expressão de incerteza e as condicionantes metodológicas são fundamentais.

Quanto ao Gemini (Google, 2024), este opera, segundo seus criadores, em um sistema no qual o modelo é incorporado a serviços de recuperação e escalonamento, que tendem a privilegiar respostas que combinam geração com base no modelo estatístico com evidência recuperada por buscas. Esta abordagem, teoricamente, seria capaz de diminuir as alucinações. Entretanto, devido à natureza diversa e desconhecida da totalidade das fontes utilizadas em buscas, há potencial para inserção de informação não fundamentada em ciência no processamento da síntese de um artigo científico. Por fim, o *DeepSeek*, segundo DeepSeek-AI (2024), usa mecanismos de Misturas de Especialistas (MoE - *Mixture of Experts*), estratégias de compressão para otimização do resultado, que difere substancialmente das arquiteturas dos outros LLM testados e sua adoção crescente em contextos não-anglófonos.

Caracterizados os modelos a serem testados, faz-se necessário elucidar as especificidades que permitam analisar possíveis vieses algorítmicos que afetem a qualidade dos outputs cujo propósito é a comunicação científica.

Motta-Roth e Hendges (2010) destacam que, especialmente em gêneros como o resumo científico, é exigido que se preserve o escopo, as condicionantes metodológicas e os limites epistemológicos do texto-base. Portanto, independentemente da ferramenta utilizada e embora os *Large Language Models* (LLMs) produzam textos linguisticamente convincentes quando resumem comunicação científica, é preciso que elas se mantenham fiéis aos elementos apontados por Motta-Roth e Hendges, isentando-se de gerar ampliações, supressões ou criar seções estruturais, introduzindo informações não fidedignas ao documento original.

A desconexão entre a coerência formal e a precisão dos fatos relatados e a característica mais marcante do fenômeno da alucinação (Maynez *et al.*, 2020; Ji *et al*., 2022).

Em tarefas de síntese[4], os autores observaram que cerca de 70% dos resumos elaborados em frases únicas apresentavam alucinações intrínsecas ou extrínsecas. Ao investigarem a origem do desse tipo de problema, Du e colaboradores (2023) apontam que as alucinações decorrem de falhas em capacidades-chave, como memorização do senso comum e raciocínio relacional seguimento de instruções, que geram afirmações ilógicas.

Somado a isso, Peters e Chin-Yee (2025) identificam a generalização como um viés estrutural específico dos LLMs na produção de resumos científicos. Segundo eles, mesmo sob instruções precisas, os modelos tendem a gerar conclusões mais amplas do que as presentes nas informações dos artigos originais. Elas foram divididas em três categorias: (a) generalização genérica, com formulação de resultados como verdades gerais e sem restrições qualificadoras; (b) generalização temporal, com afirmações descritas de forma atemporal ou estável; e (c) generalização orientada à ação, com prescrições práticas e recomendações normativas não justificadas pelo texto-fonte.

No contexto do português brasileiro (PTBR), o problema da fidedignidade pode ser agravado pela assimetria representacional nos dados de treinamento, conforme discutido por Bender *e colaboradores* (2021). Os modelos disponíveis são majoritariamente treinados com conteúdos em língua inglesa e refletem a hegemonia cultural e econômica do Norte Global (Benjamin, 2019). Essa centralização tecnológica perpetua-se porque apenas grandes corporações internacionais detêm a infraestrutura indispensável para o funcionamento desses sistemas. Segundo Saura Garcia (2024), esse tipo de monopólio caracteriza um cenário de "datafeudalismo", em que a dominação industrial clássica é substituída pelo controle psicológico e comportamental por meio de direcionamento algorítmico. Tal viés pode criar lacunas de desempenho em idiomas menos representados quantitativamente.

Em decorrência da escassez de dados contextuais localizados, os LLMs operam por mecanismos de transferência interlinguística (Yoo *et al*., 2025). Em tarefas complexas, o modelo realiza uma tradução interna implícita para o inglês, que atua como uma representação latente principal. Quando esse processo falha, o desempenho declina, e as manifestações visíveis são o *code-switching* (alternância de código com inserções morfológicas e lexicais de duas ou mais línguas); colagens sintáticas (quando a estrutura gramatical e a ordem das palavras de uma língua estrangeira são traduzidas e aplicadas literalmente em outra língua) e imprecisões conceituais que desconfiguram o jargão científico específico (local).

Tais distorções provocadas por limites tecnológicos gera efeitos que transcendem falhas técnicas. Ruha Benjamin (2019) denuncia que este tipo de processo pode pode automatizar e ocultar o racismo, a discriminação e o colonialismo digital sob uma falsa capa de objetividade técnica. Benjamin denominou este fenômeno de "Novo Código Jim" (*The New Jim Code*), também foi discutido por Cantarini (2023). Como os dados de treinamento refletem desigualdades históricas, os algoritmos tendem a reproduzir e amplificar tais disparidades. No caso da síntese de textos em português brasileiro, esse viés pode se manifestar, por exemplo, na diluição de marcadores de cautela metodológica, apagamento de saberes locais, simplificação de especificidades contextuais e substituição de elaborações de prudência condicional por sentenças generalizantes e prescritivas.

[4] Elaboração textual de um resumo de texto longo.

3. METODOLOGIA

Esta pesquisa foi desenvolvida em abordagem mista (quanti-qualitativa) visando uma análise aprofundada da compatibilidade entre resumos de artigos científicos originais e suas versões geradas por três IAG, com foco na identificação e qualificação da fidedignidade com o conteúdo do artigo, existência de possíveis vieses de generalização e de inobservância da ética em pesquisa. Os relatos de pesquisa selecionados são de pesquisa de campo na área de Educação. O estudo foi desenhado para expandir investigação anterior sobre a acurácia de LLMs na resumo de textos científicos desenvolvido por Peters e Chin-Yee (2025).

Adotou-se delineamento inspirado em experimentos comparativos[5] (Jackson; Cox, 2013) no qual os três IAG geraram nove resumos de cinco artigos científicos selecionados aleatoriamente de um banco de 50 artigos, sob três tipos distintos de prompt. Cada resumo gerado foi avaliado por três especialistas, que o compararam com o resumo original e com o conteúdo integral do artigo. Ampliando a abordagem de Peters e Chin-Yee (2025), que quantificaram, eles próprios, a ocorrência de três tipos de generalização (genérica, de tempo presente e orientadoras de ação) atribuindo zero para ausência e um para presença de cada uma, nesta pesquisa optou-se pela aplicação de uma escala de 5 pontos, apresentada na Figura 1.

Figura 1 - Escala ordinal criada para avaliação dos resumos gerados pelas IAG

Fonte: Banco de dados da pesquisa.

A escala foi utilizada como resposta às seguintes questões: (a1) Até que ponto a sequência estrutural do resumo da IAG é compatível com a estrutura do resumo original? (a2) A linguagem, coesão e ausência/presença de ambiguidade do resumo da IAG são compatíveis com o resumo original? (a3) O resumo da IAG é compatível com os elementos centrais da comunicação original sem omissão ou expansão indevida? (a4) O resumo gerado preserva a direção (para mais/para menos, presença/ausência de efeito) e a magnitude do resultado, sendo compatível com o que é reportado no artigo? (b1) O resumo gerado é compatível no que se refere à ausência de contradições/discriminações em dados, números, métodos ou resultados do original? (b2) O resumo da IAG manteve informações que podem ser verificadas no artigo original, sem deturpar significados? (b3) A IAG mantém resultados de um estudo específico sem modificar para afirmações universais/generalizadas incompatíveis com o texto original? (b4) A IA é incompatível com resultados descritivos modificando-os para recomendações de ação ausentes no artigo? (b5) O resumo da IA introduz linguagem estigmatizante ou generalizações sobre grupos sendo incompatível com o texto original? (b6) O resumo da IA é compatível com o tom emocional do original, sem apelo indevido nem neutralização excessiva?

[5] Consiste em submeter unidades experimentais a diferentes tratamentos ou condições e comparar sistematicamente as respostas observadas.

As questões compõem duas dimensões: (QC) qualidade de conteúdo, composta pelos itens a1 a a4, que avalia a qualidade do conteúdo gerado pela IAG em comparação com o padrão do resumo original; (VR) isenção de vieses e desvios em relação à responsabilidade ética da comunicação científica, itens b1 a b6. Para completar a análise, solicitou-se dos avaliadores que elaborassem justificativas textuais para toda pontuação inferior a quatro. Estes dados permitiram aprofundar a compreensão das pontuações atribuídas, especificamente quanto às incompatibilidades que prejudicam a precisão desejada para síntese dos artigos. Além disso, considerou-se como medida geral a Percepção de Qualidade (PQ) composta pelo escore obtido com a totalização dos 10 itens do instrumento.

A extração do resultado quantitativo da avaliação se deu pela soma das pontuações referentes ao QC, VR e PQ e transformação em escores proporcionalizados (resultado entre zero e um). No caso de VR, considerando a escala ordinal adotada, quanto maior a pontuação menor seria a incidência de vieses, foi aplicada a inversão de escore (um menos o escore proporcional) para que, quanto maior a pontuação proporcionalizada, maior a representação de inadequação identificada, facilitando a interpretação.

3.1 Procedimentos

Para compor corpus textual do experimento, foi solicitado que dez pesquisadores da área de educação e de informática aplicada à educação enviassem, cada um, cinco artigos escolhidos segundo as próprias percepções de relatos de pesquisa de boa qualidade. Os artigos enviados estão publicados em periódicos revisados por pares e com classificação no sistema Qualis-CAPES (extratos A ou B). Foi organizada uma pasta com os artigos enviados, nomeados com códigos. Foi criada uma planilha associando os códigos com os títulos e autores visando posterior identificação para análise de resultados. Foi realizado, então, um sorteio considerando-se apenas os códigos dos arquivos, que resultou na obtenção dos cinco textos que constam na relação a seguir e que foram utilizados na experimentação: (001-PIL) A institucionalização da Educação a Distância no Brasil e a formação de professores: subsídios para se pensar o futuro da modalidade (Chaquime; Pareschi, 2025); (002-VER) Predição de Sucesso Acadêmico de Estudantes: Uma Análise sobre a Demanda por uma Abordagem baseada em Transfer Learning (De los Reyes et al., 2019); (003-AIN) A inscrição do sujeito na escrita acadêmica numa perspectiva dialógica (Braga; Pereira, 2016); (004-NIA) Atendimento Educacional Especializado: Reflexões sobre a Demanda de Alunos Matriculados e a Oferta de Salas de Recursos Multifuncionais na Rede Municipal de Manaus-AM1 (Santos et al,. (2017) e (005-OPA) O papel do outro na constituição da profissionalidade de professoras iniciantes (André et al.,2017).

Os artigos sorteados foram preparados por meio de um *script* elaborado em Python[6], o qual extraiu o conteúdo que se inicia na seção Introdução e termina na Conclusão (ou Considerações Finais). O conteúdo extraído foi armazenado no formato texto simples (TXT). Este procedimento garantiu que os modelos testados não recebessem o resumo original já elaborado pelos autores.

Organizado o corpus, o arquivo TXT de cada artigo foi submetido isoladamente às três IAG (*ChatGPT, Gemini e DeepSeek).* Em cada submissão foi incluído um tipo de prompt diferente, conforme apresentados no Quadro 1. Esta variação nas condições de *prompting*

[6] linguagem de programação de característica interpretativa e propósito geral, utilizada por sua sintaxe simples e ampla aplicabilidade em desenvolvimento de software, análise de dados, automação e inteligência artificial. Mais informações no site da linguagem: https://www.python.org/.

visou garantir que diferentes formas de uso, comumente adotadas por usuários, fossem avaliadas em relação à geração de vieses, observação relevante para a identificação de eventuais imprecisões geradas pelas IAG.

Quadro 1 - Tipos de Prompt Utilizados

Tipo de Prompt	Característica Principal	Objetivo
Base	Leia o artigo fornecido e produza um título, um resumo e palavras-chave para ele. Comece escrevendo o título que representa o foco do artigo. Em seguida, escreva o resumo do artigo com as seguintes regras: 1) Inclua o objetivo do estudo, metodologia, principais resultados e conclusões ou considerações finais em um único parágrafo com, no máximo, 250 palavras. 2) Não use listas, tópicos, símbolos, equações, fórmulas ou diagramas. O texto deve ser contínuo e conciso. Depois, pule uma linha, escreva o rótulo: "Palavras-chave:" e, em seguida, de 3 a 5 termos que representem os conceitos principais do artigo. Separe os termos por ponto e vírgula e termine com ponto.	Definir regras rígidas para a tarefa e estilo textual do output. (White et al., 2023)
Roleplay (Persona Pattern)	Atribuição de persona (Especialista em Redação) e inclusão dela na estrutura do tipo Base: Você é um 'Especialista em Redação Acadêmica' com experiência em síntese de literatura científica multidisciplinar. Sua missão é transformar artigos complexos em resumos técnicos normatizados. [seguido dos comandos do prompt Base]	Induzir comportamento especializado na execução da tarefa. (White et al., 2023)
Emotion	Inclusão de estímulos emocionais no tipo Roleplay: Por favor, me ajude [...] Conto com sua competência pois isto é muito importante para a minha carreira).	Modular o tom da resposta e potencializar o desempenho. (Li et. al, 2023)

Fonte: Elaborado pelos autores

A submissão simultânea aos três modelos de IAG e a obtenção dos *outputs* se deu pela aplicação de outro *script* Python, cuja função foi abrir sessões com cada servidor da IAG, enviar o prompt seguido do conteúdo do artigo no formato TXT, receber o resultado, gravá-lo em uma planilha eletrônica com identificação do código do artigo, tipo de prompt e modelo de IAG, e encerrar a sessão. Este procedimento visou garantir que cada IAG não memorizasse a tarefa anterior realizada, fato que poderia afetar resultados dos *outputs*. O script foi executado três vezes (uma para cada tipo de prompt) com o conteúdo extraído de cada um dos cinco artigos, simultaneamente para as três IAG testadas. Esta operação resultou em nove outputs (resumos gerados) que foram organizados em uma matriz de três tipos de prompt por três modelos de IAG, para cada um dos cinco artigos testados.

Em seguida, os 45 resumos gerados neste processo foram preparados para avaliação por especialistas[7], por meio da organização uma planilha online contendo, por linha: um código para o avaliador, o código do artigo utilizado no teste, o resumo original, um link para

[7] Participaram da avaliação 15 pesquisadores experientes (doutores), bem como doutorandos e um mestre vinculados à 2 programas de pós-graduação em educação, todos com artigos publicados em periódicos com classificação no indicador Qualis-Capes.

acesso à íntegra, em PDF, do referido artigo, um dos resumos gerados por IAG, a sequência de questões para atribuição de pontuação, segundo a escala apresentada na Figura 1, e um campo para justificativa. Foram aplicados filtros para acesso exclusivo de cada avaliador, restringindo a análise somente aos itens destinados a ele. Procedeu-se, então, à avaliação pelos especialistas. Ao todo, foram avaliados 135[8] pares de resumo original e resumo gerado por IAG.

4. RESULTADOS E DISCUSSÃO

Os resultados e a discussão foram organizados em três eixos analíticos que se complementam, quais sejam: a validação do instrumento utilizado para avaliação dos resumos gerados pelas IAG; o desempenho das IAG, consideradas as técnicas de *prompt* aplicadas, bem como a análise quanti-qualitativa dos vieses observados.

4.1 Validação Psicométrica do Instrumento de avaliação

A consistência interna da escala utilizada na avaliação foi verificada pelo coeficiente Alfa de Cronbach ($\alpha = 0,86$), que indicou homogeneidade adequada entre os itens. Também foram investigadas evidências de validade baseadas na estrutura interna da escala. Para isso, realizou-se uma análise de componentes principais (ACP) dos dez itens do instrumento, utilizando-se o método de extração por componentes principais e rotação Varimax. O resultado indicou dois componentes com autovalores superiores a 1, responsáveis conjuntamente por 65,0% da variância total dos dados. Na solução não rotacionada, o Componente 1 apresentou autovalor de 4,602, explicando 46,0% da variância, enquanto o Componente 2 apresentou autovalor de 1,894, correspondendo a 18,9%. De forma acumulada, ambos explicaram 65,0% da variância. Após a rotação, observou-se uma redistribuição da variância entre os componentes: o Componente 1 passou a explicar 38,4%, associado à soma das cargas ao quadrado de 3,840, e o Componente 2 passou a explicar 26,6%, com soma das cargas ao quadrado de 2,657. Apesar dessa redistribuição, a variância total explicada permaneceu inalterada em 65,0%, indicando que a solução rotacionada apenas favoreceu uma interpretação mais equilibrada e parcimoniosa das dimensões subjacentes. A solução obtida se mostrou compatível com a estrutura bidimensional prevista durante a elaboração dos itens, inspirada no estudo de Peters e Chin-Yee (2025).

A matriz rotacionada consolidou a estrutura prevista, sugerindo que as avaliações dos especialistas se estruturaram em torno dos dois eixos principais: a identificação de possíveis vieses nas representações produzidas (VR) e a avaliação da qualidade do conteúdo dos resumos (QC). Os itens foram distribuídos em duas dimensões. A primeira reuniu os itens b1 a b6. a segunda foi constituída pelos itens a1 a a4, tal como concebido na elaboração do instrumento. A adequação da matriz de correlações que dá consistência à análise de componentes principais foi confirmada pelo índice de Kaiser-Meyer-Olkin (KMO = 0,79), indicando boa adequação amostral, e pelo Teste de Esfericidade de Bartlett, que se revelou estatisticamente significativo ($\chi^2 = 764,57$; $gl = 45$; $p < 0,001$). As medidas de adequação amostral individuais (MSA) variaram entre 0,58 (item A3) e 0,90 (item B3), demonstrando a adequação dos dados à análise. Em conjunto com a consistência interna, estes resultados

[8] O total de 135 comparações foi resultante da geração de três tipos de resumos por três modelos de IAG, para cada um dos cinco artigos sorteados, que foram analisadas por três especialistas (3x3x5x3=135)

foram considerados adequados para validar as análises desta pesquisa. Concluída a validação do instrumento, procedeu-se à análise do desempenho das IAG por artigo.

4.2 Desempenho das IAG por artigo resumido

Para a análise do desempenho por artigo partiu-se da obtenção dos resultados descritivos e do conjunto de pontos atribuídos. A Tabela 1 apresenta os escores das medidas obtidas. A mediana foi utilizada devido à natureza ordinal da escala aplicada (Jamieson, 2004).

Os resultados do teste de Shapiro–Wilk indicaram que a maioria das variáveis não apresentou distribuição normal. A falta de normalidade observada reflete a concentração das avaliações em regiões específicas da escala: medianas elevadas para QC, medianas reduzidas para VR (indicador cuja menor pontuação representa menor viés) e medianas elevadas para PQ, que integra a qualidade de conteúdo e ausência de viés. Assim, a não normalidade é compatível com a observação geral de boa conformidade dos resumos gerados com os originais.

Tabela 1 - Resultados estatísticos da avaliação por artigo analisado

Vari-ável	Código Artigo	Mediana	Teste de Shapiro-Wilk	P-value do Shapiro-Wilk	Mínimo	Máximo
QC	001-PIL	0,950	0,865	,002	0,750	1,000
QC	002-VER	0,850	0,921	,042	0,650	0,950
QC	003-AIN	0,900	0,784	< ,001	0,750	1,000
QC	004-NIA	0,900	0,820	< ,001	0,300	1,000
QC	005-OPA	0,950	0,805	< ,001	0,850	1,000
VR	001-PIL	0,067	0,829	< ,001	0,000	0,200
VR	002-VER	0,100	0,915	,031	0,000	0,233
VR	003-AIN	0,167	0,780	< ,001	0,033	0,767
VR	004-NIA	0,100	0,796	< ,001	0,000	0,667
VR	005-OPA	0,067	0,914	,028	0,000	0,133
PQ	001-PIL	0,940	0,918	,035	0,820	1,000
PQ	002-VER	0,840	0,927	,060	0,760	0,980
PQ	003-AIN	0,860	0,824	< ,001	0,540	0,980
PQ	004-NIA	0,920	0,838	< ,001	0,320	1,000
PQ	005-OPA	0,940	0,907	,020	0,900	0,980

Fonte: Dados da pesquisa processados por estatística descritiva no aplicativo JASP

Os escores elevados de PQ, cujas medianas variaram entre 0,84 e 0,94, indicam que, na visão dos avaliadores, os resumos gerados pelas IAG atenderam, na maioria, aos critérios de qualidade de conteúdo e apresentaram baixa incidência de vieses. Também se realizou o teste de Kruskal-Wallis, que identificou diferenças estatisticamente significativas entre os tipos de artigo, (H(4)=24,040, p<0,001). Para confirmar as diferenças específicas aplicou-se o teste post hoc de Dunn, com ajuste dos valores de *p* pelo método de Holm. Foram observadas diferenças significativas entre os artigos 001-PIL e 002-VER (z = 2,986, p_{Holm} = 0,020, r_{rb} = -0,549); 001-PIL e 003-AIN (z = 3,479, p_{Holm} = 0,005, r_{rb} = -0,546); 002-VER e 005-OPA (z = -3,393, p_{Holm} = 0,006, r_{rb} = 0,694); seguido de 003-AIN e 005-OPA (z=-3,886, p_{Holm} = 0,001, r_{rb} = 0,679$). As demais comparações não apresentaram diferenças estatisticamente significativas, depois de aplicado o ajuste de Holm (p_{Holm} >0,05). Considerando que valores mais elevados da variável analisada representam melhor desempenho, a ordem descritiva dos artigos, do melhor para o pior resultado e com base nas médias dos postos, foi: 005-OPA, 001-PIL, 004-NIA, 002-VER e 003-AIN.

Considerando a necessária cautela com a significância estatística das comparações, optou-se por verificar se havia indicações qualitativas que pudessem associar as pontuações de PQ, QC e VR com a natureza específica de cada artigo, seja pela forma de tratamento dos dados ou pela comunicação da metodologia e dos resultados, visto que elas poderiam indicar possíveis efeitos nos outputs.

A análise dos artigos que obtiveram os melhores resultados (001-PIL e 005-OPA - QC elevada e VR muito baixa) demonstrou que estes apresentam fundamentação teórica evidenciada de forma adequada, clareza argumentativa e rigor no detalhamento dos procedimentos metodológicos, bem com na discussão dos resultados. Em 001-PIL, por exemplo, a discussão sobre a institucionalização da EaD é amparada por apontamentos de legislação e de relatórios oficiais, percebeu-se ausência de juízos de valor e manutenção de tom mais analítico. Tal estabilidade argumentativa, hipotetiza-se, pode ter permitido uma compreensão mais acurada dos avaliadores e, ao mesmo tempo, diminuiu a margem de construções textuais generalizantes das IAG.

Em 005-OPA, a investigação sobre professores iniciantes baseou-se em grupos de discussão com consentimento livre e esclarecido (critério de ética em pesquisa) explicitamente declarado, o texto explicitou a descrição do método utilizados e resultados que articularam a escuta dos participantes com a discussão com quadro teórico, apresentado no referencial do artigo. Assim como no anterior, considera-se que tal estrutura pode ter contribuído para reduzir a interferência de viés interpretativo quando da leitura do texto pelos avaliadores e também ter evitado imprecisões na criação do resumo pelas IAG.

Entre os cinco textos do corpus, estes apresentaram o menor número de pontuações inferiores a quatro e, consequentemente, poucas justificativas. Identificou-se, também, que houve um ganho de qualidade de conteúdo em um resumo criado por IAG, como se observa no comentário inserido pelo Avaliador E2 sobre o output produzido pelo ChatGPT 5 com prompt *Emotion*:

> Dado interessante: o resumo da IA tem uma incongruência com relação ao original, ao mencionar que os grupos de discussão são oriundos de duas universidades e uma faculdade [...]. Mas ao verificar o artigo na íntegra, constata-se que esse dado metodológico está correto conforme o corpo do manuscrito. Ou seja, a IA corrigiu um erro [omissão] do resumo original (E2).

Não obstante a aparente qualidade de conteúdo, foram identificadas algumas imprecisões conceituais que, na visão de avaliadores, resultaram em pontuação 3 para o item B2 nos três tipos de prompt nas submissões às IAG, conforme registro do Avaliador E5 ao analisar o resumo produzido pelo DeepSeek com prompt *Emotion*:

> A Nota 3 atribuída no item B2 refere-se à inserção do termo "representações" como foco de análise. Trata-se de um termo que requer cuidado teórico-metodológico para o seu uso, pois, no campo da pesquisa em educação, este pode denotar conceito e método de análise específicos e que não foram tratados nesta pesquisa (E5).

O artigo 002-VER apresentou escore pouco menor que os anteriores (PQ = 0,920) e apresenta características de comunicação científica marcadamente distintas deles. A temática, apesar de ser de interesse educacional, tem característica técnico-computacional. O texto do artigo apresenta algumas imprecisões detectadas pelos avaliadores, como no exemplo das justificativas do Avaliador E12, ao pontuar itens do resumo gerado pelo ChatGPT com prompt *Emotion*:

> A2: [...] A metodologia inclui a informação de aproximadamente 1 milhão de registros de log utilizados, fato descrito no artigo mas excluído do resumo original. A informação no texto é aproximada (mais de 1 milhão) e no resumo da IA se atribuiu a mesma imprecisão, algo que rebaixa evidência de validade do método (E12).

Observa-se que há deficiências comunicacionais no artigo que foram apropriadas pela IAG ao elaborar o resumo. Por outro lado, identificou-se que há penalizações com justificativas compatíveis com falhas relacionadas a Vieses produzidos pelas IAG, no caso, omissões. O avaliador E6, ao comparar o resumo original com o gerado pelo DeepSeek e prompt *Role-play* declarou:

> [...] o resumo da IA é conciso e direto, não apresentando contextualização necessária aos Leitores (o que é feito no resumo original). Sobre [...] as palavras-chave sugeridos pela IA, não são condizentes com o original (E6).

Quanto à VR, o avaliador E3, ao verificar o output do ChatGPT com prompt *Base*, identificou falhas relativas à extrapolação de conclusões e omissão de informação relativa a dados oriundos de pessoas (viés de privacidade):

> [...] IA apresenta o termo "conclui" não presente no original e traz a conclusão ligada a compatibilidade prévia não presente no original. Resumo da IA não cita [que a pesquisa trabalhou com] os dados de mais de 3000 alunos (logs) (E3).

Os escores mais baixos (menores que 90% de conformidade com a pontuação total) aparecem em 003-AIN e 004-NIA, sobretudo na VR, o que é relevante por se tratar do indicador de vieses. Em 003-AIN, os resultados produzidos concentram-se na coleta com participação de um único aluno, selecionado justamente por apresentar maior dificuldade na escrita acadêmica (“esse aluno apresentou um maior grau de dificuldade em relação aos demais colegas”). A exposição de imagens do texto manuscrito do participante e a densa

interpretação discursiva, ainda que respaldada em Bakhtin, podem ter enviesado a produção do output, visto que as imagens não foram extraídas para submissão (limitação do método adotado). Por exemplo, o avaliador E8, ao analisar o output do CHatGPT com prompt *Base* destacou:

> Nota 3: Maioria dos elementos presentes, mas ao menos um relevante ausente ou expandido.[...] IA indicou o tipo de metodologia (não mencionada no original); especificou gênero analisado (resenha) que, no original, foi indicado apenas como "produção textual decorrente da uma atividade avaliativa"; resumo da IA não indicou o critério de seleção do corpus, que foi apontado no original; [...] O resumo da IA faz generalização sobre letramento acadêmico, mesmo que a pesquisa tenha sido baseada em um estudo de caso específico. Tom geral preservado, mas ao menos uma limitação relevante foi omitida. (E8).

Em 004-NIA, o estudo é quantitativo sobre salas de recursos multifuncionais e, além de dados de fonte secundária discutidos por estatística descritiva, os autores fazem extrapolações a partir de premissas arbitrárias (e.g. 40 alunos por SRM) com justificativa pouco fundamentada nos próprios dados coletados (“supomos calcular que a capacidade máxima … seria … quarenta alunos por semana”). A partir desta constatação, hipotetizou-se que esta forma de apresentação de resultados quantitativos, somada à ausência de explicitação de limitações na metodologia (recorte da pesquisa) pode ter contribuído para a atribuição de menor pontuação. O avaliador E9, ao analisar o *output* do DeepSeek gerado com prompt *Emotion* destacou: “[...] ela [a IA] oscila (omite) sobre como é colocado no texto original o período (anos) da coleta e dos resultados”. Ao analisar o *output* da mesma IAG obtido com o prompt *Role-play*, E9 justifica: “Neste caso a IA volta a inferir distorções relevantes e generalizantes sobre elementos importantes do texto e faz conclusões que podem ser precipitadas ou gerais [...]. O texto parece começar a seguir o tom [do] original, mas se perde a partir das considerações metodológicas”. Outro avaliador (E1) também identificou distorções relevantes geradas pelo DeepSeek com prompt *Role-play*:

> Nota 3 em B1 porque um dado apresentado é impreciso, ainda que o texto [do resumo], no geral, esteja correto e com base nos texto original [...] tem um acréscimo de dado relevante [da IAG] e que não foi apresentado no texto original (E1).

Em suma, os resultados de comparação dos escores por artigo sugerem que a forma de organização e apresentação das informações metodológicas e dos resultados no texto do relato de pesquisa podem interferir na qualidade da geração do output, independentemente da IAG ou tipo de prompt utilizado. Esta constatação aponta para a necessidade de supervisão humana especializada para validação das informações sintetizadas. Ou seja, há impossibilidade de substituição da leitura, pelo pesquisador, de um relato de investigação por uma automação de elaboração de resumos, por exemplo, na atividade de revisão da literatura. Observou-se, nesta pequena amostra, que estudos que se basearam em casos específicos ou que empregam interpretações de dados quantitativos sem a explicitação clara dos limites para generalização, tiveram a transposição destas falhas para os resumos gerados por IA, resultando em falsas generalizações ou extrapolações. Em outras palavras, há indícios de que textos de comunicação científica imprecisos possam gerar resumos

automatizados ainda mais imprecisos. Além da comparação entre modelos, investigou-se o efeito do tipo de prompt sobre a qualidade dos resumos gerados.

4.3 Percepção de Qualidade analisada por IAG e por tipo de prompt

Concluída a análise baseada no conteúdo de cada artigo, realizou-se a análise do desempenho global, focada na comparação dos resultados por modelo de IAG (Tabela 2). As medianas demonstram bom desempenho geral na tarefa solicitada, o que não significa, necessariamente, precisão comunicacional do ponto de vista científico.

Tabela 2 - Resultados descritivos de avaliação por modelo de IAG

Variável	IAG	Mediana	Teste de Shapiro-Wilk	P-value do Shapiro-Wilk	Mínimo	Máximo
QC	ChatGPT 5.0	0,90	0,76	< ,001	0,30	1,00
QC	Gemini 3.0 Flash	0,90	0,88	< ,001	0,55	1,00
QC	DeepSeek V3	0,90	0,79	< ,001	0,50	1,00
VR	ChatGPT 5.0	0,07	0,69	< ,001	0,00	0,77
VR	Gemini 3.0 Flash	0,13	0,75	< ,001	0,00	0,70
VR	DeepSeek V3	0,07	0,70	< ,001	0,00	0,70
PQ	ChatGPT 5.0	0,92	0,76	< ,001	0,32	1,00
PQ	Gemini 3.0 Flash	0,90	0,81	< ,001	0,50	1,00
PQ	DeepSeek V3	0,94	0,80	< ,001	0,44	1,00

Fonte: Dados da pesquisa processados por estatística descritiva no aplicativo JASP.

Em termos descritivos, o escore de QC, que capturou a fidelidade à estrutura do artigo [problema, objetivo, metodologia, resultados, conclusão], clareza e limite de abrangência, indicou que as três IAG tiveram o mesmo desempenho (Md = 0,90). Todas foram percebidas pelo conjunto dos avaliadores como capazes de produzir compatibilidade elevada dos resumos gerados com os originais, adotados como referência.

No escore de VR, cuja medição indica o nível de incidência de alucinações ou falhas éticas na comunicação, o DeepSeek e o ChatGPT apresentaram desempenho ligeiramente melhor (Md = 0,07) que o Gemini (Md = 0,13). Observa-se que a diferença é pequena (6%) para sustentar que ele é menos indicado para a tarefa. A interpretação mais relevante desta quantificação é a comparação entre valores mínimos e máximos, pois indicam casos extremos, nos quais a percepção de vieses ou desvios foram significativos. Estes resultados corroboram a percepção inicial, obtida na análise qualitativa das avaliações por artigo, que sugere a impossibilidade de atribuição de confiabilidade nas sínteses automatizadas, a ponto de se dispensar a supervisão humana.

A variável PQ, que acumula a qualidade do conteúdo aliada à desejável inexistência de vieses, indicou uma leve desvantagem do Gemini (Md = 0,90), em relação ao ChatGPT

(Md=0,92) e ao DeepSeek V3 (Md=0,94). A diferença também é muito pequena e o resultado geral elevado (superior a 90% de compatibilidade em relação ao escore máximo), colocando os modelos de IAG como equivalentes na tarefa testada. Isto mostra que, mesmo com tecnologias de treinamento diferentes, o desempenho geral dos outputs pode ser considerado sintaticamente coerente, guardadas as restrições qualitativas relacionadas aos vieses.

O teste de Kruskal-Wallis para PQ confirmou a equivalência, pois não se identificou diferenças estatisticamente significativas decorrentes do modelo de IAG ($H(2) = 1,60$, $p = 0,448$). As comparações post hoc de Dunn também não identificaram diferenças significativas entre os modelos de IA (todos os valores ajustados foram superiores a 0,05). As correlações bisseriais por postos foram pequenas (entre −0,09 e 0,15). A partir destas constatações confirmou-se ausência de destaque para uma das tecnologias, apesar de suas diferenças estruturais no treinamento.

Ao lado disso, não obstante as limitações qualitativas identificadas, o resultado da percepção da qualidade foi próximo entre as três IAG testadas, independentemente da arquitetura subjacente (seja o modelo comercial otimizado por RLHF da OpenAI, o ecossistema integrado do Gemini, ou a arquitetura MoE do DeepSeek). Tal observação está de acordo com a premissa de Wei et al. (2022) sobre as "habilidades emergentes" dos LLMs, que aponta o aumento exponencial de parâmetros como viabilidade técnica para que esses sistemas gerem textos sintaticamente aceitáveis. Contudo, como adverte Kaufman (2022), essa capacidade sintática aceitável reflete um salto estatístico-computacional e não uma compreensão epistêmica dos textos gerados.

Para aprofundar a compreensão sobre esta percepção de vieses na qualidade, analisou-se possível efeito do tipo de prompt utilizado nos resultados. A Tabela 3 apresenta a estatística descritiva dos escore de PQ por modelo de IAG e tipo de prompt utilizado.

Tabela 3. Resultados descritivos de avaliação por modelo de IAG e Tipo de Prompt

Var	IAG	Tipo de Prompt	Mediana	Teste de Shapiro-Wilk	P-value do Shapiro-Wilk	Mínimo	Máximo
PQ	ChatGPT	Base	0,90	0,70	,002	0,48	0,98
PQ	ChatGPT	Emotion	0,94	0,70	< ,001	0,32	1,00
PQ	ChatGPT	RolePlay	0,92	0,77	,002	0,54	1,00
PQ	DeepSeek	Base	0,90	0,85	,015	0,54	1,00
PQ	DeepSeek	Emotion	0,94	0,84	,012	0,60	1,00
PQ	DeepSeek	Roleplay	0,94	0,73	< ,001	0,44	0,98
PQ	Gemini	Base	0,90	0,81	,005	0,50	0,98
PQ	Gemini	Emotion	0,88	0,81	,005	0,56	1,00
PQ	Gemini	Roleplay	0,92	0,79	,003	0,56	1,00

Fonte: Dados da pesquisa processados por estatística descritiva no aplicativo JASP.

Na Tabela 3 é possível observar que os resultados medianos de PQ são elevados (Md=0,88 ou mais) para os três tipos de prompts utilizados, sugerindo percepção geral muito boa, fato esperado tendo em vista os resultados das análises por artigo e por modelo de IAG. Para os três modelos de IAG testados, o tipo Base recebeu os menores escores (Md=0,90). O tipo RolePlay, que incorporou a definição de uma persona, gerou um ganho mediano de 2% para ChatGPT e Gemini, bem como de 4% para o DeepSeek. No caso do tipo Emotion, que incorporou estímulos emocionais aos comandos do tipo Roleplay, houve comportamento distinto, com perda de desempenho de 4% (Md=0,88) no caso de Gemini, em comparação com o resultado do Roleplay (Md=0,92).

Entretanto, o teste de Kruskal–Wallis para PQ não indicou diferenças estatisticamente significativas provocadas pelo tipo de prompt ($H(8) = 2,506$, $p = 0,961$). Assim, confirmou-se que as diferenças observadas na Tabela 3 não são suficientes para a determinação de que incrementos no desempenho tenham sido motivados por variações nos comandos dos prompts. Tal constatação se contrapõe às proposições de White et al. (2023) aos resultados de Li et al. (2023), neste estudo. Na perspectiva conceitual proposta por White et al. (2023), a adoção de enquadramentos estruturados, como o Persona Pattern (Roleplay), estabelece um contexto de atuação que orienta o comportamento do modelo, resultando em respostas com maior coerência e profundidade em comparação a instruções básicas ou desestruturadas. A hierarquia de desempenho foi corroborada por resultados de Li et al. (2023), os quais demonstraram que a incorporação de estímulos emocionais e psicológicos (EmotionPrompt) provocou melhoria de 11% nas métricas de desempenho, veracidade e responsabilidade em tarefas de geração de texto. Os ganhos observados no presente estudo são muito modestos em relação aos observados por Li et al.

Retomando a percepção de qualidade, é importante destacar que as medianas mínimas e máximas observadas tanto na tabela 2 quanto na 3, indicam que nem todos resumos sintéticos foram fidedignos em relação ao que os relatos de pesquisa descrevem, apenar dos escores medianos elevados (entre 0,88 e 0,94). Para este tipo de tarefa testado neste estudo, mesmo pequenos desvios de precisão podem gerar falhas com potencial para induzir leitores a erros não admissíveis na produção de conhecimento científico.

Tal observação também encontra evidências em extratos de conteúdos presentes nos resumos originais em comparação com *outputs*, conforme apresentado no Quadro 2, cujos exemplos ilustram a metáfora dos "papagaios estocásticos" de Bender et al. (2021).

Quadro 2 - Comparação de conteúdos dos resumos originais com outputs gerados

Original (004-NIA)	**Output ChatGPT (Prompt Base)**
Objetivo: Discutir as ofertas das Salas de Recursos Multifuncionais (SRM) em Manaus e sua conformidade legal. **Metodologia:** Pesquisa quantitativa descritiva, levantamento documental federal e dados do INEP/SEMED (2009-2015). **Resultados:** Oferta insuficiente de SRM frente ao aumento de matrículas; déficit estimado de 30 salas.	**Objetivo:** Analisar a adequação das SRM às políticas de inclusão em Manaus. **Metodologia:** Abordagem quantitativa baseada em Censo Escolar e SEMED (período 2014-2016). **Resultados:** Déficit de 40% (30 novas salas); necessidade de replanejamento e expansão imediata.
Original (002-VER)	**Output ChatGPT (Prompt Emotion)**
Objetivo: Análise empírica para verificar indícios de diferenças entre dados provenientes de contextos educacionais	**Objetivo:** Verificar empiricamente a existência de divergências em distribuições de dados educacionais utilizadas em tarefas de predição

Original (004-NIA)	Output ChatGPT (Prompt Base)
distintos na tarefa de predição de insucesso acadêmico de estudantes. **Metodologia:** Emprega-se dados de logs de mais de 3.000 estudantes de ensino superior na modalidade EAD. A metodologia adotada é baseada na própria abordagem de classificação supervisionada, **Resultados:** Ainda que o cenário de dados envolva atividades comuns a estudantes de uma mesma disciplina, os experimentos indicam uma acurácia de até 83%na separação de dados provenientes de períodos letivos distintos. Embora empíricos, os resultados indicam uma direção similar àquela apontada por outros trabalhos [...]voltados a prevenção de insucessos acadêmicos.	de insucesso acadêmico em cursos superiores a distância. **Metodologia:** Abordagem quantitativa baseada na classificação supervisionada, utilizando algoritmos de Árvores de Decisão e Support Vector Machine aplicados a mais de um milhão de registros de interação de estudantes com o ambiente Moodle. **Resultados:** Os resultados indicam acurácia superior a 65% na distinção de contextos em três das quatro disciplinas analisadas, com destaque para a disciplina de Lógica, que apresentou valores acima de 80%, [...]. Conclui-se que a detecção prévia de covariate shift pode aprimorar a acurácia de predições e fundamentar o uso de técnicas de transfer learning [...]

Fonte: Base de dados da pesquisa

Observa-se que, ao prever o próximo *token* mais provável, a IAG parece descartar marcadores de incerteza, substituindo-os por argumentos mais assertivos e diretos (e.g.: "discutir a oferta" se transformou em “analisar a adequação”, "oferta insuficiente [...] déficit estimado" se transformaram em “Déficit de 40% [... e] necessidade de replanejamento e expansão imediata”). Além disso, a fidelidade factual (subgrupo de itens com a letra B no instrumento) sofreu adulterações. O melhor exemplo é o resumo automatizado de 002-VER (Quadro 2), no qual a IAG adotou abordagem distinta do original para informar sobre tamanho da amostra, distorcendo a comunicação científica original (e.g. “Emprega-se dados de logs de mais de 3.000 estudantes de ensino superior” se transformou em “[...] algoritmos [...] aplicados a mais de um milhão de registros de interação de estudantes com o ambiente Moodle”).

Como se pôde perceber, a ilusão de que as IAG são competentes colide com o fato de que os LLM não possuem consciência ou compromisso com a verdade. Utilizam cálculos estatísticos baseados em probabilidades para combinar padrões de texto utilizados no treinamento e prever a próxima palavra mais provável. Por gerarem textos fluentes e formalmente coerentes, os usuários menos experientes no seu uso podem atribuir-lhes características humanas (fenômeno da antropomorfização), elevando o risco de perda de fidedignidade e confiabilidade na circulação do conhecimento.

5. CONCLUSÃO

A comparação do desempenho das IAG (ChatGPT 5.0, Gemini 3.0 Flash e DeepSeek V3) testados especificamente na tarefa de produzir resumos de relatórios de pesquisa publicados em português brasileiro indicou que os três modelos apresentam bom desempenho sintático, no geral, bem como paridade no nível de desenvolvimento tecnológico de modo que, nas condições adotadas, não se observou diferença relevante em favor de um deles. Destaca-se, entretanto, que o desenho metodológico adotado não autoriza generalização de resultados para outros contextos. Portanto, o fato de não se identificarem diferenças relevantes e de se ter observado percepção de boa qualidade não são indicativos

robustos de que se pode utilizar qualquer uma para se obter resultado satisfatório na tarefa de síntese de textos científicos ou relatos de pesquisa.

A mensuração da percepção de qualidade dos modelos apresentou medianas entre 84% e 94% do escore total, indicando que a maioria das avaliações apontou níveis elevados de compatibilidade com os originais. O teste de Kruskal-Wallis identificou a inexistência de diferenças significativas entre os modelos ou tipos de prompt, confirmando a proximidade de desempenhos aparentemente satisfatórios dos modelos testados. Contudo, a análise qualitativa dos artigos, das justificativas dos avaliadores, a comparação entre resumos originais e outputs de casos nos quais pontuação foi menor que 90% do escore total identificou tendências à generalização ou à supressão de marcadores de incerteza presentes nos textos originais.

Tais achados apontam para fragilidades na fidedignidade de parte dos resumos produzidos, com potencial de prejudicar a exatidão da comunicação científica pretendida pelos autores dos artigos. A reflexão necessária, neste caso, se refere à estabilidade e confiabilidade dos outputs. Se em 10% dos casos o output falha ao comunicar com precisão o objetivo, a metodologia e ou principal resultado de uma pesquisa, isto é suficiente para que o pesquisador dispense conferência minuciosa das sínteses de uma série relatórios de pesquisa que precisam resumidos?

Os resultados obtidos neste estudo também sugeriram uma possibilidade de existência de vieses oriundos do processo de transferência interlinguística (Yoo et al., 2025), que consiste na execução de uma tradução opaca para o inglês durante o processamento no LLM. Como os modelos operam majoritariamente em inglês (Benjamin, 2019; Saura Garcia, 2024), o processamento de comandos complexos em português demanda que o modelo transfira essas instruções para seu contexto de trabalho e, neste processo, nuances específicas e emocionais engendradas na comunicação português podem ser “diluídas”, interferindo no *output* resultante. Novos estudos poderão confirmar esta possibilidade.

Os resultados obtidos também indicam a necessidade de realização de nova etapa deste estudo incluindo modelo de IAG treinado no idioma português. Segundo apontam Pires et al. (2023), modelos especializados na língua portuguesa podem ser capazes de capturar a textura narrativa e o jargão específico da ciência produzida nesta língua.

Conclui-se, portanto, que a “habilidade” de síntese da comunicação científica observada nos modelos de IAG testados é predominantemente sintática, carecendo de maior rigor epistêmico. O fenômeno da supergeneralização (a transformação de achados limitados em verdades universais) identificada por Peters e Chin-Yee (2025) e a supressão ou extrapolação de marcadores de incerteza foram confirmados e representam riscos para a integridade acadêmica da comunicação processada pelos modelos de IAG.

Esses desvios, muitas vezes ocultos sob uma redação fluida e convincente, podem induzir pesquisadores, notadamente os que estão em processo de em formação, a interpretações errôneas e identificação de evidências distorcidas.

Em suma, embora as IAG sejam apoios disponíveis para a produção intelectual, os resultados deste estudo alertam para a impossibilidade de dispensa da supervisão humana especializada. A leitura atenta de relatórios de pesquisa e a validação crítica dos outputs permanecem como etapas indispensáveis do fazer científico.

REFERÊNCIAS

BENDER, Emily M.; GEBRU, Timnit; MCMILLAN-MAJOR, Angelina; SHMITCHELL, Shmargaret. On the dangers of stochastic parrots: can language models be too big? In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2021, Virtual Event, Canada. **Proceedings [...]**. New York: Association for Computing Machinery, 2021. p. 610–623. Disponível em: https://doi.org/10.1145/3442188.3445922. Acesso em: 19 maio 2026.

BENJAMIN, Ruha. **Race after technology: abolitionist tools for the New Jim Code**. Cambridge: Polity, 2019.

BROWN, Tom B.; MANN, Benjamin; RYDER, Nick; SUBBIAH, Melanie; KAPLAN, Jared D.; DHARIWAL, Prafulla; NEELAKANTAN, Arvind; SHYAM, Pranav; SASTRY, Girish; ASKELL, Amanda; AGARWAL, Sandhini; HERBERT-VOSS, Ariel; KRUEGER, Gretchen; HENIGHAN, Tom; CHILD, Rewon; RAMESH, Aditya; ZIEGLER, Daniel M.; WU, Jeffrey; WINTER, Clemens; HESSE, Christopher; CHEN, Mark; SIGLER, Eric; LITWIN, Mateusz; GRAY, Scott; CHESS, Benjamin; CLARK, Jack; BERNER, Christopher; MCCANDLISH, Sam; RADFORD, Alec; SUTSKEVER, Ilya; AMODEI, Dario. Language models are few-shot learners. In: **Advances in Neural Information Processing Systems 33**. Red Hook: Curran Associates, 2020. p. 1877–1901. Disponível em: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.htm. Acesso em: 23 ago. 2026.

CANTARINI, Paola. **Filosofia da inteligência artificial com base nos valores construcionistas do 'homo poieticus'**. Rio de Janeiro: Lumen Juris, 2023.

COMITÊ GESTOR DA INTERNET NO BRASIL. **Pesquisa sobre o uso das tecnologias de informação e comunicação nas escolas brasileiras: TIC Educação 2024**. São Paulo: CGI.br, 2025. Disponível em: https://cetic.br/pt/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nas-escolas-brasileiras-tic-educacao-2024/ . Acesso em: 23 ago. 2026.

CRAWFORD, Joseph; COWLING, Michael; ALLEN, Kelly-Ann. Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI). **Journal of University Teaching & Learning Practice**, [S. l.], v. 20, n. 3, art. 2, 2023. DOI: https://doi.org/10.53761/1.20.3.02. Disponível em: https://doi.org/10.53761/1.20.3.02 . Acesso em: 23 ago. 2026.

DEEPSEEK-AI. **DeepSeek-V3 technical report**. arXiv, 2024. arXiv:2412.19437. Disponível em: https://arxiv.org/abs/2412.19437. Acesso em: 23 ago. 2026.

DU, Li; WANG, Yequan; XING, Xingrun; YA, Yiqun; LI, Xiang; JIANG, Xin; FANG, Xuezhi. **Quantifying and attributing the hallucination of large language models via association analysis**. arXiv, 2023. arXiv:2309.05217. Disponível em: https://arxiv.org/abs/2309.05217. Acesso em: 18 maio 2026.

FUNDAÇÃO ITAÚ; DATAFOLHA. **Consumo e uso de inteligência artificial no Brasil**. São Paulo: Observatório Fundação Itaú, 2025. Disponível em: https://www.fundacaoitau.org.br/observatorio/biblioteca/pesquisa-consumo-e-uso-da-inteligencia-artificial-no-brasil. Acesso em: 22 ago. 2026.

GOOGLE. **Gemini 1.5: our next-generation model, plus a new 1.5 Flash model**. Google DeepMind, 2024. Disponível em: https://deepmind.google/technologies/gemini/. Acesso em: 10 ago. 2026.

JACKSON, Michelle; COX, D. R. The principles of experimental design and their application in sociology. **Annual Review of Sociology**, Palo Alto, v. 39, p. 27–49, 2013. DOI: https://doi.org/10.1146/annurev-soc-071811-145443. Disponível em: https://doi.org/10.1146/annurev-soc-071811-145443 Acesso em: 23 ago. 2026.

JAMIESON, Susan. Likert scales: how to (ab)use them. **Medical Education**, [S. l.], v. 38, n. 12, p. 1217–1218, 2004. DOI: https://doi.org/10.1111/j.1365-2929.2004.02012.x. Disponível em: https://doi.org/10.1111/j.1365-2929.2004.02012.x. Acesso em: 27 mar. 2026.

JI, Ziwei; LEE, Nayeon; FRIESKE, Rita; YU, Tiezheng; SU, Dan; XU, Yan; ISHII, Etsuko; BANG, Yejin; CHEN, Delong; DAI, Wenliang; CHAN, Ho Shu; MADOTTO, Andrea; FUNG, Pascale. Survey of hallucination in natural language generation. **ACM Computing Surveys**, New York, v. 55, n. 12, art. 248, p. 1–38, 2023. DOI: https://doi.org/10.1145/3571730. Disponível em: https://doi.org/10.1145/3571730. Acesso em: 23 ago. 2026.

JURAFSKY, Daniel; MARTIN, James H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models**. 3. ed. [S. l.: s. n.], 2026. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 23 ago. 2026.

KAUFMAN, Dora. **Desmistificando a inteligência artificial**. 2. ed. rev. e ampl. Belo Horizonte: Autêntica, 2022.

LI, Cheng; WANG, Jindong; ZHANG, Yixuan; ZHU, Kaijie; HOU, Wenxin; LIAN, Jianxun; LUO, Fang; YANG, Qiang; XIE, Xing. **Large language models understand and can be enhanced by emotional stimuli**. arXiv, 2023. arXiv:2307.11760. Disponível em: https://arxiv.org/abs/2307.11760. Acesso em: 30 jul. 2026.

LIMA, Jéssica Caroline Rodrigues de; FELIPPE, Maíra Longhinotti. Integridade acadêmica na era do ChatGPT: desafios éticos e as novas fronteiras da inovação. **Revista Brasileira da Educação Profissional e Tecnológica**, [S. l.], v. 3, n. 25, p. 1–22, e17803, 2025. DOI: https://doi.org/10.15628/rbept.2025.17803. Disponível em: https://doi.org/10.15628/rbept.2025.17803. Acesso em: 27 jul. 2026.

MAYNEZ, Joshua; NARAYAN, Shashi; BOHNET, Bernd; MCDONALD, Ryan. On faithfulness and factuality in abstractive summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58., 2020, Online. **Proceedings [...]**. Online: Association for Computational Linguistics, 2020. p. 1906–1919. DOI: https://doi.org/10.18653/v1/2020.acl-main.173. Disponível em: https://aclanthology.org/2020.acl-main.173/. Acesso em: 23 ago. 2026.

MOTTA-ROTH, Désirée; HENDGES, Graciela Rabuske. **Produção textual na universidade**. São Paulo: Parábola Editorial, 2010.

OPENAI. **GPT-4 technical report**. arXiv, 2023. arXiv:2303.08774. Disponível em: https://arxiv.org/abs/2303.08774. Acesso em: 14 ago. 2026.

PETERS, Uwe; CHIN-YEE, Benjamin. Generalization bias in large language model summarization of scientific research. **Royal Society Open Science**, London, v. 12, n. 4, 241776, 2025. DOI: https://doi.org/10.1098/rsos.241776. Disponível em: https://doi.org/10.1098/rsos.241776. Acesso em: 23 ago. 2026.

PIRES, Ramon; ABONIZIO, Hugo; ALMEIDA, Thales Sales; NOGUEIRA, Rodrigo. Sabiá: Portuguese large language models. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 12., 2023, Belo Horizonte. **Intelligent Systems: proceedings, Part III**. Cham: Springer, 2023. p. 226–240. (Lecture Notes in Computer Science, v. 14197). DOI: https://doi.org/10.1007/978-3-031-45392-2_15. Disponível em: https://doi.org/10.1007/978-3-031-45392-2_15. Acesso em: 23 ago. 2026.

SAMPAIO, Rafael Cardoso; NICOLÁS, Maria Alejandra; JUNQUILHO, Tainá Aguiar; SILVA, Luiz Rogério Lopes; FREITAS, Christiana Soares de; TELLES, Márcio; TEIXEIRA, João Senna; ESCÓSSIA, Fernanda da; SANTOS, Luiza Carolina dos. ChatGPT e outras IAs transformarão a pesquisa científica: reflexões sobre seus usos. **Revista de Sociologia e Política**, Curitiba, v. 32, e008, 2024. DOI: https://doi.org/10.1590/1678-98732432e008. Disponível em: https://doi.org/10.1590/1678-98732432e008 Acesso em: 27 jul. 2026.

SAURA GARCÍA, Carlos. Datafeudalism: The Domination of Modern Societies by Big Tech Companies. **Philosophy & Technology**, v. 37, n. 3, p. 90, set. 2024.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia. Attention is all you need. In: **Advances in Neural Information Processing Systems 30**. Red Hook: Curran Associates, 2017. p. 5998–6008. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Acesso em: 23 ago. 2026.

WEI, Jason; TAY, Yi; BOMMASANI, Rishi; RAFFEL, Colin; ZOPH, Barret; BORGEAUD, Sebastian; YOGATAMA, Dani; BOSMA, Maarten; ZHOU, Denny; METZLER, Donald; CHI, Ed H.; HASHIMOTO, Tatsunori; VINYALS, Oriol; LIANG, Percy; DEAN, Jeff; FEDUS, William. Emergent abilities of large language models. **Transactions on Machine Learning Research**, [S. l.], 2022. Disponível em: https://openreview.net/forum?id=yzkSU5zdwD. Acesso em: 28 jul. 2026.

WHITE, Jules; FU, Quchen; HAYS, Sam; SANDBORN, Michael; OLEA, Carlos; GILBERT, Henry; ELNASHAR, Ashraf; SPENCER-SMITH, Jesse; SCHMIDT, Douglas C. A prompt pattern catalog to enhance prompt engineering with ChatGPT. In: CONFERENCE ON PATTERN LANGUAGES OF PROGRAMS, 30., 2023, Allerton Park, Monticello, Illinois. **Proceedings [...]**. [S. l.]: Association for Computing Machinery, 2023. DOI: https://doi.org/10.64346/PLoP2023p05. Disponível em: https://www.plopcon.org/proceedings/plop/2023/05.html. Acesso em: 23 ago. 2026.

XIAO, Ping; CHEN, Yuanyuan; BAO, Weining. **Waiting, banning, and embracing: an empirical analysis of adapting policies for generative AI in higher education**. arXiv, 2023. arXiv:2305.18617. Disponível em: https://arxiv.org/abs/2305.18617. Acesso em: 31 jul. 2026.

YOO, Haneul; JIN, Jiho; CHO, Kyunghyun; OH, Alice. **Code-switching in-context learning for cross-lingual transfer of large language models**. arXiv, 2025. arXiv:2510.05678. Disponível em: https://arxiv.org/abs/2510.05678. Acesso em: 20 maio 2026.

APÊNCICE

Referência bibliográfica dos artigos utilizados para geração de resumos pelos IAG

(001-PIL) CHAQUIME, Luciane Penteado; PARESCHI, Claudinei Zagui. A institucionalização da Educação a Distância no Brasil e a formação de professores:

subsídios para se pensar o futuro da modalidade. **Dialogia**, São Paulo, n. 54, p. 1-19, e28459, Ed. Esp., maio/ago. 2025. DOI: https://doi.org/10.5585/54.2025.28459.

(002-VER): DE LOS REYES, Daniel A. Guimarães; THOMAS, Everton André; ROSA, Lilian Landvoigt da; GAVIÃO NETO, Wilson P. Atendimento educacional especializado em Manaus-AM. **Revista Brasileira de Informática na Educação – RBIE**, v. 27, n. 1, p. 1-24, 2019. https://doi.org/10.5753/RBIE.2019.27.01.01.

(003-AIN): BRAGA, Sandro; PEREIRA, Rodrigo Acosta. A inscrição do sujeito na escrita acadêmica numa perspectiva dialógica. **Fórum Linguístico**, Florianópolis, v. 13, n. 3, p. 1506-1524, jul./set. 2016. DOI: http://dx.doi.org/10.5007/1984-8412.2016v13n3p1506.

(004-NIA) SANTOS, João Otacílio Libardoni dos et al. Atendimento educacional especializado: Reflexões sobre a demanda de alunos matriculados e a oferta de Salas de Recursos Multifuncionais na Rede Municipal de Manaus-AM. **Revista Brasileira de Educação Especial**, Marília, v. 23, n. 3, p. 409-422, jul.-set., 2017. DOI: http://dx.doi.org/10.1590/S1413-65382317000300007.

(005-OPA) ANDRÉ, Marli; CALIL, Ana Maria Gimenes Corrêa; MARTINS, Francine de Paulo; PEREIRA, Marli Amélia Lucas. O papel do outro na constituição da profissionalidade de professoras iniciantes. **Revista Eletrônica de Educação**, v. 11, n. 2, p. 505-520, jun./ago., 2017. DOI: http://dx.doi.org/10.14244/198271992231.

Declaração de contribuição dos autores (Taxonomia CRediT)

Ronei Ximenes Martins: Conceptualization, Project administration, Methodology, Formal analysis, Writing – original draft.
Licenciado em Matemática (FAFI-UNIS/MG) e Doutor em Psicologia pela (USF). Professor Associado da Universidade Federal de Lavras (UFLA). Professor Permanente do Programa de Pós-Graduação em Educação e Coordenador do Núcleo de Estudos em IA aplicada ‘a Educação da UFLA. rxmartins62@gmail.com

Maria Kamylla e Silva Xavier: Methodology, Formal analysis, Writing – review & editing.
Graduada em Física (UFRN). Mestra em Ensino de Ciências Naturais e Matemática (UFRN). Doutora em Educação (UFPB). Professora Adjunta da Universidade Federal de Campina Grande (UFCG). Professora permanente do Programa de Pós-graduação em Educação da UFCG. kamylla.xavier@professor.ufcg.edu.br

Bruno Amarante Couto Rezende: Conceptualization, Writing – original draft.
Graduado em Sistemas de Informação (UNIPAC). Mestre em Educação (UNIVAS). Doutorando em Educação (UFLA). Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Campus Varginha. brunorezende@cefetmg.br

Nicole de Santana Gomes Garcia: Data curation, Writing – original draft.
Graduada em Letras/Português (UFLA). Mestra em Educação (UFLA) Doutoranda em Educação (UFLA). Professora da Universidade do Estado de Minas Gerais (UEMG), unidade Campanha/MG. nicole.gomes@uemg.br

Patricia Peixoto Carneiro Viegas: Writing – original draft, Writing – review & editing.

Graduada em Pedagogia (Universidade Estácio de Sá). Mestra em Educação (UFLA) Doutoranda em Educação (UFLA). Professora do Centro Universitário Presidente Tancredo de Almeida Neves (UNIPTAN). patricia.viegas1@estudante.ufla.br

Francine de Paulo Martins Lima: Formal analysis, Writing – review & editing.
Graduada em Pedagogia pela (UMC/SP). Doutora e Mestre em Educação (PUC-SP). Professora Adjunta da área de Didática e estágios da Universidade Federal de Lavras - UFLA. Líder do Grupo de Pesquisa sobre Formação docente e práticas pedagógicas - FORPEDI (CNPq/UFLA). francine.lima@ufla.br

Declaração de conflito de interesse

Os autores declaram que não há conflito de interesse.

Declaração de disponibilidade de dados da pesquisa

O conjunto de dados de apoio aos resultados deste estudo foi disponibilizado no formato de painel do Google Data Studio com acesso livre para leitura e pode acessado em *https://datastudio.google.com/reporting/1f5199a2-784b-4280-b08b-076003f09004]*.

Declaração de uso de IA

- Conforme descrito no seção de metodologia, os modelos CHatGPT 5.0, Gemini 3.0 Flash e DeepSeek V3 foram utilizados para geração dos outputs analisados.

- Na elaboração do texto do artigo, houve utilização da ferramenta de Inteligência Artificial NotebookLM, na organização das fontes bibliográficas em núcleos temáticos e para confirmação de citações indiretas ou paráfrases no momento de revisão da escrita. Na revisão final da ortografia utilizou-se a ferramenta Sabiá 4.0 para sugestão de correção de eventuais erros de digitação ou ajuste de estilo da escrita, bem como para conferência da paridade entre citações autor-data no texto e a descrição completa das obras na seção de referência.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.