BRUNO, Adriana Rocha. Itens Digitais na avaliação: experiências com as inteligências artificiais generativas na área de Língua Portuguesa. *TECCOGS* — Revista Digital de Tecnologias Cognitivas, n. 32, p. 118-135, 2025

Recebido em: 3 dez. 2024
Aprovado em: 23 dez. 2024

dx.doi.org/10.23925/1984-3585.2025i32p118-135

Lincensed under CC BY 4.0

Itens Digitais na avaliação: experiências com as inteligências artificiais generativas na área de Língua Portuguesa[1]

Adriana Rocha Bruno[2]

Resumo: Este artigo apresenta resultados de uma investigação realizada entre 2023 e 2025 sobre a utilização de inteligências artificiais generativas (IAGen) na elaboração de Itens Digitais (ID) para avaliações em larga escala no Brasil, com fomento do CAED/UFJF. O estudo compreendeu um levantamento bibliográfico nacional e internacional, e experimentações empíricas com Perplexity e Gemini para a produção de itens em Língua Portuguesa. Foram produzidos 42 itens de LP, testando níveis de complexidade e estratégias de coerência textual. Os resultados apontam que as IAGen ampliam a velocidade e a diversidade na elaboração de itens, mas apresentam limitações quanto a coerência pedagógica, rigor científico e risco de vieses. Do ponto de vista metodológico, foram estabelecidas estratégias que envolveram: (i) geração inicial com Perplexity; (ii) estruturação e refinamento com Gemini; e (iii) validação docente crítica por equipe de pesquisadores especialistas da área de língua portuguesa. A pesquisa evidencia implicações éticas no uso das IAGens, alertando para a necessidade de regulação responsável, curadoria docente e metodologias híbridas. A análise mais aprofundada de LP evidenciou que há necessidade de exame e anuência de especialistas da área e que as IAGens possuem conhecimento limitado acerca de temas singulares, usando, por exemplo, sinônimos inadequados, desconsiderando regras específicas e distratores convergentes, bem como não atendendo à plausibilidade para a construção de itens. Conclui-se que as IAGens são muito úteis e podem auxiliar sobremaneira no desenvolvimento de itens, mas não substituem a mediação humana, e se configuram como interlocutoras críticas para repensar a avaliação em larga escala na educação básica brasileira.

Palavras-chave: Inteligência artificial generativa; Itens Digitais; avaliação na educação.

---

1 Este artigo é produto de pesquisa financiada pela Fundação CAED-UFJF (2023-2025).

2 Doutora e Mestre em Educação: Currículo pela PUC-SP, com Pós-doutorado em educação pelo Instituto de Educação da Universidade de Lisboa-PT. Licenciada em Pedagogia, foi professora da Educação Básica por 17 anos. É Professora Titular do Departamento de Didática e permanente do PPGEdu da UNIRIO. É autora do livro: *Formação de professores na cultura digital: aprendizagens do adulto, educação aberta, emoções e docências,* publicado pela Editora da UFBA, com a versão em e-book aberta no site da editora. Email: adriana.bruno@unirio.br. ORCID: https://orcid.org/0000-0002-5646-8919.

Digital Items in assessment: experiences with generative artificial intelligence in the portuguese language area

Abstract: This article presents the results of research conducted between 2023 and 2025 on the use of generative artificial intelligence (GAI) in the development of Digital Items (DI) for large-scale assessments in Brazil, with funding from CAED/UFJF. The study comprised a national and international bibliographic survey and empirical experiments with Perplexity and Gemini for the production of items in Portuguese. Forty-two LP items were produced, testing levels of complexity and textual coherence strategies. The results indicate that IAGen increases the speed and diversity in the development of items, but has limitations in terms of pedagogical coherence, scientific rigour, and risk of bias. From a methodological point of view, strategies were established that involved: (i) initial generation with Perplexity; (ii) structuring and refinement with Gemini; and (iii) critical teacher validation by a team of researchers specialising in the Portuguese language. The research highlights ethical implications in the use of IAGs, warning of the need for responsible regulation, teacher curation, and hybrid methodologies. A more in-depth analysis of LP showed that there is a need for examination and approval by specialists in the field and that IAGens have limited knowledge about specific topics, using, for example, inappropriate synonyms, disregarding specific rules and convergent distractors, as well as not meeting the plausibility requirements for item construction. It is concluded that IAGs are very useful and can greatly assist in the development of items, but they do not replace human mediation and are critical interlocutors for rethinking large-scale assessment in Brazilian basic education.

Keywords: Generative artificial intelligence; Digital Items; assessment in education.

Introdução – Avaliações: interatividade como potência para mudanças

As tecnologias digitais no campo da educação, como desdobramentos da cultura digital, nos convidam a rever concepções e práticas pedagógicas, pois alteram formas de acesso e de construção de conhecimentos, e mudam os modos de relação entre as pessoas. Os processos de aprendizagem e de avaliação não estão apartados deste cenário, e os dispositivos digitais e em rede, mais do que máquinas e ou recursos tecnológicos, são instrumentos culturais e de aprendizagem. É neste contexto que, estima-se, os processos de ensino, de aprendizagem e os de avaliações possam a ser integrados aos dispositivos tecnológicos disponíveis.

Em pesquisas nacionais e internacionais realizadas no período de 2018 a 2022 (Bruno, 2025), os relatórios e artigos científicos acusavam que, historicamente, as avaliações em larga escala têm sido desenvolvidas e aplicadas no formato impresso, com predominância de itens de múltipla escolha e/ou produção de texto (questões dissertativas). Os estudos mostraram que, mesmo com o avanço das tecnologias digitais ao longo do século XXI, o desenvolvimento de novos formatos de itens de avaliação ainda se espelha predominantemente em avaliações impressas e são somente adaptados para o formato digital, sem (ou com pouquíssima) alteração.

A avaliação educacional, tema ainda nevrálgico, como sabemos, deveria ser compreendida, fundamentalmente, em sua conexão com as aprendizagens. Isso porque, para nós, não deve haver avaliações desconectadas dos processos de ensino e de aprendizagem. Trazer tais proposições nos faz pensar: é possível desenvolver avaliação sem fins de reprovação? É possível avaliação sem classificar? Compreendemos que a avaliação como processo ainda está em construção. Sobre as avaliações externas, também denominadas 'de larga escala', Fernandes nos alerta:

> [...] Essa cultura se dá no interior das escolas e fora delas, quando a partir das avaliações externas realizadas pelas redes municipais e/ou estaduais, como exemplo do que vem acontecendo agora, são feitos os famosos rankings ou são premiadas escolas e professores. Embora, no discurso educacional contemporâneo, as avaliações ganhem destaque pelo fato de serem fundamentais para a garantia da qualidade das escolas, o que vemos são os usos dos resultados das avaliações para outros fins. (Fernandes, 2015, p. 22)

Pluralizamos tanto avaliação quanto educação – avaliações e educações –, porque não há, em nosso entender, unicidade quando se trata destes temas.

Segundo dados das pesquisas desenvolvidas anteriormente pelas autoras Liamara Scortegagna e Adriana Rocha Bruno (2022), há uma relação intensa entre as concepções de aprendizagem, as culturas escolares e as ideias de avaliação. Fernandes (2022) caminha em direção similar. As formas de avaliar refletem a cultura, os pensamentos, as ideias, as concepções e as políticas (ou suas respostas e desdobramentos). Portanto, as avaliações em larga escala – ainda conservadoras e alicerçadas, em sua maioria, em ranqueamentos, na classificação e, principalmente, na exclusão –, utilizam-se dos propósitos de culturas escolares/educacionais para fomentar o medo, as disputas, as diferenças socioculturais e tantos outros elementos. Isso ocorre em detrimento da valorização das aprendizagens e da utilização das avaliações como percursos transformativos, nos quais o saber e o não saber sejam partes do processo de aprender. Infelizmente, ainda mais distantes do que ocorre com as avaliações da aprendizagem, as em larga escala, por serem em massa (e não para e com as pessoas) não realizam seus possíveis objetivos, a saber: avaliar para promover mais aprendizagens.

Não queremos, com isso, dizer que as avaliações da aprendizagem superaram os obstáculos para se firmarem como processos de aprender, mas que talvez um caminho interessante para as avaliações em larga escala esteja na sua aproximação com as da aprendizagem, quando formativas.

Destarte, em acordo com Thais Saçço (2015), os dados (índices) produzidos por boa parte das políticas públicas nacionais com as avaliações em larga escala, para acompanhar o desenvolvimento das escolas, fazem-se por meio de ranqueamentos, no lugar de diagnósticos e de proposições para melhorar ou solucionar os problemas atinentes à educação.

Nesse cenário, que destoa de um dos grandes ensinamentos freirianos, "denunciar e anunciar" (Freire, 1979), a avaliação em larga escala poderia se prestar a identificar de que forma as políticas públicas estão de fato cumprindo seu papel de fomentar, financiar, sustentar, alicerçar, fundamentar e criar as condições necessárias para que alunos e professores desenvolvam aprendizagens e docências, e para que a escola desenvolva seu papel pedagógico, psicossocial e cultural.

A avaliação em larga escala no Brasil se constitui como um instrumento estratégico de política pública educacional, fornecendo dados sobre o desempenho de estudantes, muitas vezes sem considerar os distintos contextos socioeconômicos e geográficos do país.

Todavia, a elaboração de itens avaliativos de qualidade permanece como um desafio substantivo. Nesse cenário, emerge a inteligência artifi-

cial generativa (IAGen) como ferramenta potencial de apoio ao processo de design de avaliações digitais interativas. Este artigo apresenta os resultados de uma pesquisa desenvolvida entre 2023 e 2025 que investigou as possibilidades e limites da IAGen para a produção de itens avaliativos em avaliação de larga escala na disciplina de Língua Portuguesa. A partir de uma abordagem teórico-metodológica que articulou dimensões técnicas, pedagógicas e éticas, buscou-se responder: de que formas as inteligências artificiais generativas podem integrar/favorecer o desenvolvimento de itens digitais avaliativos?

IAGen, educação e avaliação: dados bibliográficos

A literatura recente sobre inteligência artificial generativa destaca seu potencial para transformar práticas de avaliação educacional tanto de aprendizagem quanto em larga escala, promovendo personalização, automação na geração de itens e análise de grandes volumes de dados.

O levantamento bibliográfico realizado, nacional e internacional, focou em publicações sobre inteligência artificial generativa, educação e avaliação em larga escala, utilizando bases de dados como Scielo, ERIC e Google Scholar. A pesquisa identificou o uso de diferentes verbetes para as buscas, como "generative artificial intelligence" e "digitally based assessments".

O período de 2023 a 2025 se caracterizou pelo crescimento exponencial de publicações sobre inteligência artificial generativa aplicada à educação, especialmente centradas no ChatGPT e seus impactos nas práticas pedagógicas (Arruda, 2024). Entretanto, observou-se disparidade notável entre a vasta produção sobre IA generativa em contextos educacionais amplos e a ainda tímida produção sobre aplicações específicas em avaliação em larga escala e desenvolvimento de itens digitais avaliativos (Tonbuloglu, 2023).

Estudos nacionais evidenciaram o potencial transformador da IAGen nas práticas docentes. Arruda (2024) discute a IA generativa no contexto da transformação do trabalho docente, enquanto Silva e Sá (2024) analisam sua aplicação ao ensino inclusivo. Loiola *et al.* (2024) propõem reflexões sobre o impacto da IAGen nos processos de ensino e de aprendizagem de língua portuguesa na educação básica, ressaltando preocupações com autoria e diversidade narrativa. Fonseca e Campiglia (2024) apresentam o framework APREND.AI, defendendo que a integração ética da IA generativa deve considerar autoria e diversidade cultural.

No âmbito internacional, Kaldaras *et al.* (2024) abordam o desenvolvimento de avaliações válidas na era da IA generativa, enfatizando que novas formas de articulação multimodal transcendem a mera digitalização do impresso. Giannakos *et al.* (2024) advertem, contudo, sobre os riscos de dependência cognitiva e respostas apenas plausíveis. Xia *et al.* (2024) e Noroozi *et al.* (2024) analisam como a IAGen está transformando a avaliação, ampliando possibilidades de feedback adaptativo e desafiando critérios de validade tradicionalmente aceitos.

O uso da IAGen na educação envolve mudanças no papel docente, na validade dos instrumentos e nas demandas por novas competências digitais entre educadores e estudantes. Estudos nacionais e internacionais ressaltam tanto benefícios – como eficiência, inclusão e *feedback* automatizado – quanto desafios éticos e técnicos relacionados a questões algorítmicas, privacidade e integridade acadêmica.

Os dados apresentados nos levantamentos bibliográficos indicam um cenário de crescimento do tema, com estudos que abordam o uso da IAGen para personalização do aprendizado e novos métodos de avaliação. No entanto, também alertam para riscos como a subjetividade na avaliação de resultados criativos e o viés ético. A literatura aponta que a avaliação com as IAGens deve ser feita por meio de métricas híbridas que combinem análises quantitativas e qualitativas. Os pesquisadores alertam também para a necessidade de diretrizes éticas no uso da IAGen em pesquisa, visando proteger a privacidade de dados.

A integração de IAGen em contextos avaliativos levanta questões éticas e sociais prementes. Conforme destaca a UNESCO (2024), a IAGen tende a restringir narrativas plurais ao reforçar pontos de vista dominantes, resultando em homogeneização do conhecimento que limita o pensamento pluralista e criativo, reduzindo a pluralidade e comprometendo perspectivas decoloniais e inclusivas. Essa padronização de respostas e conformidade representam risco à autonomia intelectual e ao desenvolvimento do pensamento crítico estudantil.

Estudos realizados por Lima e Serrano (2024) identificam ainda categorias críticas para avaliação de IA em educação: limitações técnicas, viés ético e impacto na aprendizagem. Preocupações com racismo algorítmico, associação automática de armas a figuras racializadas, e produção de *fake news* mais convincentes que narrativas humanas delineiam um cenário de complexidade e urgência regulatória, como aponta o relatório de pesquisa de Bruno (2025a).

As IAs generativas intensificam tensionamentos acerca dos processos decoloniais e contracoloniais, exigindo escolhas conscientes sobre os caminhos da educação. Ideias e fundamentos de autores como Paulo Freire (1979), Bell Hooks (2017), Franz Fanon (2008) e Nêgo Bispo (Santos, 2023) desconstroem visões conservadoras alicerçadas em práticas coloniais.

Metodologia e produção de dados: desenho da pesquisa

Esta pesquisa se configurou como um estudo exploratório de cunho qualitativo, desenvolvido em fases articuladas e progressivas entre agosto de 2023 e setembro de 2025. A questão norteadora foi: de que formas as inteligências artificiais generativas podem integrar/favorecer o desenvolvimento de itens digitais avaliativos?

Primeira etapa (2023-2024): levantamento bibliográfico e mapeamento tecnológico

Foi realizado um levantamento bibliográfico sistemático em bases nacionais (SciELO, Repositórios UFRJ) e internacionais (Scopus, ERIC, SSRN) com os seguintes descritores: "inteligência artificial generativa", "educação", "avaliação", "avaliação em larga escala" e "itens digitais". O período focalizou publicações de 2023 a 2025. Identificaram-se 18 obras diretamente relacionadas a IAGen e 10 sobre itens digitais em educação/avaliação, apresentados anteriormente.

Paralelamente, recursos tecnológicos gratuitos para IAGen foram mapeados e testados, tais como: TextFX, Gradescope, Nuance, Perplexity, WriteSonic, ChatSonic, Gemini(Bard), ChatGPT, Gamma, ChatPDF, NotebookLM, Canvas, dentre outros. Cada plataforma foi testada em contextos simulados de produção de itens e textos, registrando-se potencialidades e fragilidades.

Segunda etapa (2024-2025): análise e produção de itens digitais avaliativos

A pesquisa avançou para a análise de itens oriundos de avaliações internacionais (PISA), utilizando a matriz criada por Bruno e Scortegagna (2022) como instrumento analítico. Foram analisadas unidades como "Leite de Vaca", "Fórum das Aves" e "Utilização de Telemóvel" (Brasil, 2020), identificando-se baixa interatividade e percursos lineares na maioria dos itens. Essa análise funcionou como uma inflexão metodológica,

reconhecendo que itens internacionais de referência ainda operavam com lógica limitada; sendo assim, a pesquisa redirecionou investimentos para a experimentação com IAGen em produção de itens com potencialidade interativa para avaliações. Esta etapa foi suspensa para dedicação intensa com as IAGens.

Terceira etapa: produção de itens digitais com IAGen e validação humana

Em articulação com equipes de pesquisadores do CAED/UFJF nas áreas de Língua Portuguesa (LP) e Matemática, produzimos aproximadamente 42 itens de LP e uma quantidade menor de itens matemáticos, utilizando plataformas como Perplexity Pro, Gemini e ChatGPT. Cada item gerado foi submetido à validação humana quanto a: pertinência pedagógica; clareza de enunciados; adequação ao perfil de avaliação em larga escala; distinção entre descritores (localização de informação *versus* produção de inferência); e plausibilidade de distratores.

Para este artigo, optou-se por trazer a análise de itens produzidos com IAGen para Língua Portuguesa, tendo em vista o aprofundamento da pesquisa nesta área.

Análise e interpretação dos resultados de Língua Portuguesa

A produção de itens de LP com IAGen evidenciou uma tensão central, segundo Begma Barbosa (2025): o potencial de escala e inovação convive com o risco de empobrecimento e homogeneização. A IAGen demonstrou elevada capacidade gerativa ao produzir grande número de itens em curto tempo, algo materialmente inviável apenas com elaboração humana. Contudo, essa velocidade foi acompanhada de fragilidades técnicas e pedagógicas significativas.

A seguir, apresentam-se dados das análises desenvolvidas por meio de itens produzidos com as IAGens, em acordo com os quatro níveis de complexidade das tarefas de leitura (Barbosa; Micarello; Ferreira, 2020). Os níveis compreendem escalas de dificuldade voltados para anos/séries da Educação Básica em Língua Portuguesa:

a) Nível 1: predominância de questões factuais e literais, em formato de múltipla escolha, com enunciados curtos e objetivos;

b) Nível 2: predominância de questões factuais e literais mais amplas, também em formato de múltipla escolha, com enunciados curtos, médios, objetivos e ampliando nível de complexidade;

c) Nível 3: questões interpretativas, que demandam do estudante a articulação entre informações do texto-base e inferências de médio alcance;

d) Nível 4: propostas mais complexas, envolvendo análise de recursos estilísticos e discursivos, próximas do que se espera em exames como ENEM.

Os pontos nevrálgicos identificados a seguir foram produzidos por Bruno (2025b) e Barbosa (2025) em meio a análises dos itens com as IAGens por meio de prompts orientados para cada nível e itens de LP:

1. Superficialidade em enunciados: muitos itens de nível 1 repetiam padrões de perguntas factuais, sem explorar dimensões interpretativas ricas. Não havia progressão cognitiva em séries longas de itens nominalmente designados para diferentes anos escolares.

2. Redundância léxico-sintática: observou-se a tendência de repetir estruturas frasais ou conceitos já apresentados no enunciado, empobrecendo a diversidade textual e oferecendo pistas indevidas para a resolução.

3. Vieses na seleção de textos-base: as plataformas privilegiaram gêneros canônicos (notícia, crônica, conto), negligenciando gêneros multimodais ou contemporâneos.

4. Notório viés discursivo em textos baseados em sensibilidade, demandando rigorosa checagem de linguagem inclusiva.

5. Ambiguidade nas alternativas: em certos casos, opções de resposta aproximavam-se demasiadamente uma da outra, comprometendo a clareza e a objetividade avaliativa. Exemplo paradigmático: itens com gabarito duplo, onde duas ou mais alternativas respondiam legitimamente à questão conforme enunciada.

6. Distinção conceitual inadequada entre descritores: a IAGen demonstrou dificuldade em distinguir entre "localização de informação explícita" e "produção de inferência". A primeira envolve recuperação de informação claramente expressa na superfície textual (por repetição, quase-repetição ou paráfrase); a segunda demanda

estabelecimento de relações entre partes não explícitas do texto. Frequentemente, a IAGen construía itens que avaliavam inferência quando solicitada localização, ou vice-versa (Barbosa, 2025).

7. Distratores não plausíveis: múltiplos itens apresentavam distratores não plausíveis, facilitando descarte imediato pelos respondentes e comprometendo a validade do instrumento. Conforme análise especializada da pesquisadora Barbosa (2025), distratores plausíveis devem apresentar informações verossímeis ao contexto, podendo conter palavras presentes no texto, mas não constituindo o gabarito.

De modo similar ao anterior, destacam-se aspectos potentes identificados nas análises dos itens produzidos com as IAGens:

1. Rapidez e escala: a capacidade generativa da IA permitiu a produção de um volume significativo de itens em curto prazo, algo inviável com equipes reduzidas de especialistas.

2. Versatilidade de gêneros: quando bem "calibrada", a IA demonstrou uma capacidade de criar itens a partir de textos de diferentes gêneros. Importante ressalva: alucinações ocorreram, incluindo invenção de textos inexistentes (como um item atribuído ao escritor Ziraldo, que não existe e foi produzido pela IAGen).

3. Estímulo à reflexão crítica em níveis avançados: alguns itens de nível 4 alcançaram análises sofisticadas de recursos retóricos e estilísticos, contribuindo para o desenvolvimento da competência leitora em níveis complexos.

4. Integração multimodal incipiente: ainda que embrionária, houve tentativas de proposição de itens com elementos visuais, sinalizando potencial de inovação.

Os dados sugerem que as IAGens apresentam limitações críticas quando os programas precisam lidar com algum conhecimento especializado – como "classes de itens" que avaliam uma mesma habilidade, distinção entre descritores específicos ou reconhecimento de "habilidades de leitura" – ou com conceitos complexos, como "plausibilidade", relacionados a aspectos técnicos da elaboração de itens (Barbosa, 2025; Bruno, 2025a).

Esses achados reforçam a necessidade de comandos que sejam capazes de lidar com a complexidade do conhecimento envolvido na produção de itens para avaliação em larga escala. Bons comandos precisam alcançar uma definição rigorosa de categorias teóricas segundo as quais sistemas

de IAGen devem operar, sobretudo quando essas categorias são específicas de um campo de conhecimento e envolvem diferenciações complexas.

Outros dois aspectos muito importantes identificados nesta pesquisa são a limitação de conhecimentos específicos que as IAGens ainda possuem no que tange à Língua Portuguesa, suas regras, critérios, especificidades gramaticais etc.; e a necessidade de produções de itens de avaliação por IAGen serem acompanhadas e validadas por especialistas da área de Língua Portuguesa.

Os vieses culturais e raciais também apareceram reiteradamente nas experimentações, especialmente na geração de imagens destinadas a compor itens digitais. A associação automática de armas a figuras racializadas, assim como a produção de imagens majoritariamente brancas quando não havia especificações étnico-raciais, evidencia a presença de racismo algorítmico entranhado nos modelos. Tal fenômeno, já amplamente documentado em diversas áreas, adquire uma dimensão particularmente grave na avaliação em larga escala, cujo compromisso fundamental é garantir equidade, justiça e ausência de discriminações estruturais.

A necessidade de regulação aparece como consequência natural desse conjunto de problematizações. Estudos recentes com docentes e estudantes indicam que normas de uso ético da IA permanecem em processo de consolidação, e que grande parte das instituições prioriza a prevenção de fraudes em detrimento da formação crítica para um uso seguro e criativo dessas tecnologias. A pesquisa sugere, portanto, que o avanço responsável do campo depende de esforços coordenados para construção de políticas claras, formação docente em letramento digital crítico, desenvolvimento de modelos linguísticos alinhados a realidades locais e garantia de soberania digital sobre bases de dados educacionais. Sem esses pilares, o risco é substituir desigualdades históricas por desigualdades algorítmicas ainda mais difíceis de mapear e enfrentar.

Ideia inovadora: calibragem de IAGen

Por meio desta pesquisa, Bruno (2025b) criou a ideia de "calibragem" de dispositivos de IAGen, entendida como um esforço sistemático de depuração dos sistemas para que reflitam a "voz" do pesquisador e respondam de forma mais contextualizada às demandas específicas da investigação.

Por calibragem, compreende-se o ato de desenvolver prompts que atendam às demandas de acordo com o desejado, ou seja, o ato de ajustar, de afinar, de treinar uma IAGen para que atenda aos comandos de forma personalizada.

A fim de ensinar a IAGen sobre o que se deseja e tentar fazer com que ela "obedeça" aos comandos ofertados pelo usuário, ensinando-a, por meio de comandos diversos, implica fazê-la corresponder ao interlocutor e compreender como poderá melhor trabalhar para atingir o que se quer/precisa.

Quando a IAGen já foi apresentada aos níveis de complexidade de 1 a 4 produzidos pela equipe de LP do CAED, foram identificadas limitações – as chamadas alucinações. Os ajustes necessários para que tais dispositivos não inventem, criem livremente e sem respeito aos parâmetros dados pelos usuários também faz parte do conceito de calibragem da IAGen.

Esse exercício pedagógico-tecnológico se revelou uma prática crítica de autoria e de resistência frente ao risco de homogeneização dos discursos gerados por máquina.

A calibragem das IAGens envolve:

1. Definição precisa de categorias teóricas com as quais o programa deve operar;

2. Apresentação de itens modelares que exemplifiquem critérios técnicos específicos;

3. Descrição detalhada de conceitos como "plausibilidade de distratores" e "distinção entre descritores";

4. Ajuste iterativo das respostas através de feedback estruturado.

A calibragem das Inteligências Artificiais Generativas (IAGen) é uma noção emergente, criada e sistematizada na pesquisa de Bruno (2025b), que ultrapassa a ideia técnica de "afinamento de modelos" (*fine tuning*) e propõe uma abordagem pedagógica, epistemológica e ética para o uso de IA em processos avaliativos.

As IAGens não produzem automaticamente itens avaliativos rigorosos, válidos e confiáveis. É necessário ensinar tais dispositivos a operarem com categorias, descritores, níveis de complexidade e argumentos justificados. Portanto, apenas após um conjunto de instruções reiteradas, progressivas e qualificadas, o sistema passa a produzir itens mais robustos e aderentes às demandas da avaliação educacional.

A IA não pensa ou produz autonomamente, mas opera por recombinação estatística de linguagem. Assim, calibrar uma IAGen é assumir a posição de autor-mediador, que orienta a produção, delimita os discursos

possíveis, cria condições para que a IA produza com rigor e que atua com responsabilidade e autoria ética.

A calibragem passa a funcionar como uma espécie de andaime cognitivo, por meio de pequenos passos graduais e suporte provisório, até que a IA consiga realizar uma tarefa mais complexa com menor intervenção.

Nessa direção, a importância da calibragem da IAGen está na possibilidade de limitar e talvez impedir que se homogenize discursos, que se reproduzam vieses algorítmicos ou mesmo que se tente apagar subjetividades. A calibragem se torna, então, uma forma de resistência crítica à automatização acrítica da educação.

Assim, calibrar uma IA não significa apenas escrever melhores prompts, mas instaurar um processo contínuo de ensinagem, diálogo e refinamento, no qual a IA passa a reconhecer padrões conceituais, metodológicos e epistemológicos específicos.

Etapas da calibragem na prática

A partir da experiência acumulada na produção de itens de Língua Portuguesa ao longo da pesquisa, foi possível sistematizar o processo de calibração das inteligências artificiais generativas em uma sequência de etapas que, embora aqui apresentadas de modo analítico, oferecem uma perspectiva contínua, dialógica e iterativa.

A primeira etapa consiste no alinhamento conceitual. Como a IAGen não diferencia autonomamente descritores, habilidades ou níveis de complexidade, tornou-se necessário alfabetizá-la nesses princípios estruturantes de avaliação em larga escala. Isso implicou ensinar-lhe o que é um item, como se define um descritor, de que modo se compreende plausibilidade, o que caracteriza os níveis de complexidade textual do 1 ao 4, e quais distinções separam habilidades como localização, inferência e inferência lexical. Trata-se de um trabalho inicial de aplicação conceitual, no qual a IAGen é integrada ao universo epistemológico e metodológico da avaliação educacional.

A etapa seguinte aborda a modelagem e a exemplificação. Identificamos que a IAGen aprende por padrões e, portanto, precisa ser exposta a modelos bem construídos para poder replicar estruturas adequadas. Nesse momento, foram apresentados itens considerados bem formulados, acompanhados de justificativas completas de gabarito, análises de erros recorrentes e comentários produzidos por especialistas da área. Esses exemplos funcionaram como matrizes de orientação discursiva e estru-

tural, permitindo que a IAGen compreendesse não apenas o formato das tarefas, mas também o raciocínio pedagógico e avaliativo subjacente.

À medida que a IAGen gera respostas, é fundamental instituir um processo sistemático de retroalimentação crítica. Nessa etapa, o papel humano é decisivo para indicar, por exemplo, que um distrator é fraco, que um gabarito apresenta dupla possibilidade, que o nível de complexidade está mal interpretado ou que o texto não se alinha ao descritor solicitado. Essa devolutiva constante produz um ciclo de aperfeiçoamento que não poder ser alcançado apenas pela máquina, pois evidencia que o desenvolvimento qualitativo depende da intervenção analítica de especialistas.

Quando a IAGen persiste em determinados erros, inicia-se uma etapa de recalibragem corretiva. Esse momento sinaliza a necessidade de reorientar, reinsistir e reensinar conceitos e procedimentos já apresentados, confirmando que a reprodução, aqui, é estratégia metodológica e não redundância. A IAGen, em sua lógica probabilística, precisa dessas retomadas sucessivas para ajustar padrões internos e reduzir recorrências internas.

Por fim, destaca-se que a calibração não produz um estado final, mas um processo contínuo. Mesmo após avanços significativos, é possível monitorar novas versões de ferramentas como Gemini, ChatGPT e Perplexity, analisar mudanças de atitude e retomar instruções sempre que os modelos apresentem regressões ou desvios. A calibração, portanto, não se encerra, mas acompanha a natureza dinâmica do IAGen e se constitui como prática permanente de mediação humana, sustentada por vigilância epistemológica e cuidado ético.

Com a calibragem, a IAGen pode passar a produzir dados com mais possibilidade de precisão, qualidade e confiabilidade. No caso desta pesquisa, os itens poderiam ser mais alinhados aos descritores, ter distratores plausíveis e não manipulativos, justificativas não superficiais, textos coerentes com o nível de complexidade e menor tendência à ambiguidade e à resposta dupla.

Mesmo calibrada, a IAGen oscila, pois sua atuação não depende unicamente do usuário, mas de seus desenvolvedores. Por isso, mais do que a colaboração crítica de usuários, é preciso que haja uma regulação das IAGens por parte do governo e das instituições.

A calibragem de IAGen funciona, portanto, como um processo de ensinagem crítica, iterativa (Santaella, 2024) e metodologicamente orientada, no qual seres humanos especialistas instruem, ajustam e refinam

as respostas de modelos generativos, de modo que estes possam operar conforme os critérios epistemológicos, linguísticos e éticos de um campo específico. Trata-se de uma mediação autoral que articula saber técnico, teórico-epistemológico e princípios éticos, reconhecendo tanto as potências quanto os limites irredutíveis das IAGens.

Considerações finais: rumo a uma educação crítica, aberta e decolonial

A pesquisa desenvolvida entre 2023 e 2025 demonstrou que inteligências artificiais generativas possuem potencial substantivo para apoiar a produção em larga escala de itens avaliativos. Sua velocidade e capacidade generativa representam ganhos reais para instituições que necessitam de grande volume de instrumentos avaliativos de qualidade.

Contudo, esse potencial permanece substancialmente limitado quando confrontado com a complexidade do conhecimento especializado em avaliação. As fragilidades críticas, a incapacidade de distinguir entre descritores específicos, a dificuldade em produzir distratores plausíveis e a tendência à homogeneização de narrativas revelam que a mediação humana qualificada permanece insubstituível.

A ideia de "calibragem" de IAGen emergiu como estratégia inovadora de resistência à homogeneização: trata-se de uma prática crítica por meio da qual pesquisadores depuram sistemas para que reflitam contextos e "vozes" específicas. Essa calibragem não é mera otimização técnica, mas se constitui como um exercício político de autoria e responsabilidade.

Compreende-se que, por meio de tal ato cotidiano, riscos como a homogeneização do conhecimento e a restrição de narrativas, citados anteriormente, possam ser atenuados. Sabemos que esses dispositivos tendem a reproduzir e reforçar perspectivas dominantes e hegemônicas em seus bancos de dados, o que pode limitar a emergência de variações discursivas e criativas, e comprometer a pluralidade epistêmica e o pensamento divergente. Em contextos avaliativos, essa homogeneização representa uma ameaça concreta à identificação de trajetórias cognitivas plurais e às diferentes formas de manifestação do raciocínio matemático.

Destarte, faz-se urgente que o investimento em pesquisas sobre IAGen como as aqui apresentadas sejam fomentadas e possam ser continuadas, particularmente envolvendo áreas das humanidades e contextos de educação para populações historicamente marginalizadas.

Dito isto, podemos pensar no desenvolvimento de protocolos éticos para o uso de IAGen em produção de avaliações, incluindo auditoria sis-

temática de vieses e garantias de diversidade narrativa. Para tanto, apresentam-se três aspectos:

1. Formação específica de avaliadores e especialistas em calibragem de dispositivos de IAGen, entendida como prática técnica e crítica;

2. Diretrizes de transparência quanto à autoria, à validação e ao uso de itens produzidos com auxílio de IAGen;

3. Integração de perspectivas decoloniais que questionem não apenas como usar a IAGen em avaliação, mas para quem, a serviço de quem e com quais finalidades.

Estima-se que esses pontos potencializem movimentos tanto para a regulação das IAGens quanto para a criação de medidas claras, éticas e plurais para o campo da avaliação.

A presente pesquisa se situa em um movimento maior de reflexão sobre o papel de tecnologias na educação brasileira. Não se trata apenas de integrar mais eficientemente máquinas em práticas avaliativas; trata-se de questionar qual educação queremos, para quem e em que termos.

A resposta que emerge desta investigação aponta para uma educação que seja aberta, critica e decolonial. Aberta para que garanta inclusão, equidade e vozes insurgentes; que não reproduza exclusões. Crítica para que questione permanentemente a quem serve determinada tecnologia, a serviço de que e de quem estão ferramentas e sistemas. Decolonial para que reconheça pluralidades epistemológicas, que seja intercultural e engajada com realidades específicas.

A IAGen pode ser uma ferramenta a serviço dessa educação, desde que rigorosamente questionada, calibrada e submetida a um escrutínio ético constante. Caso contrário, pode aprofundar desigualdades, homogeneizar pensamentos e apagar vozes. A escolha não é técnica, é profundamente política.

Referências

ARUDA, Eucidio Pimenta. Inteligência artificial generativa no contexto da transformação do trabalho docente. *Educação em Revista*, v. 40, e28495, 2024.

BARBOSA, B. T. A produção de itens para avaliação em leitura pela inteligência artificial: análise de resultados e proposição de comandos. *Relatório Interno de pesquisa*. Fundação CAED, Produto 2, 2025.

BARBOSA, B. T.; MICARELLO, H. L.; FERREIRA, R. V. Avaliações em larga escala de língua portuguesa: uma pesquisa sobre a complexidade dos textos que dão suporte a itens que avaliam a leitura. *Pesquisa e Debate em Educação*, 10(1), 1064–1081. 2020.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Brasil no PISA 2018* [recurso eletrônico]. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020.

BRUNO, Adriana Rocha. Relatório final: Itens digitais interativos na avaliação em larga escala. *Circulação Interna CAED-UFJF*. Universidade Federal do Estado do Rio de Janeiro, 2025a.

BRUNO, Adriana Rocha. Conceito de calibração de IAGen. Desenvolvido no contexto da pesquisa sobre itens digitais interativos. *Circulação Interna CAED-UFJF*, 2024-2025b.

BRUNO, Adriana Rocha, SCORTEGAGNA, Liamara. e-Val: um protótipo para análise e avaliação de níveis de interatividade em itens digitais. *Pesquisa e Debate em Educação*, Juiz de Fora: UFJF, v. 12, n. 1, p. 1-26, e36983, jan./jun. 2022. Disponivel em: https://periodicos.ufjf.br/index.php/RPDE/article/view/36983

FANON, Frantz. *Pele negra, máscaras brancas*. Salvador: EDUFBA, 2008.

FERNANDES, Claudia de Oliveira. Uma breve análise das políticas de avaliação e sua relação com a organização escolar por ciclos: resultados de pesquisa. *Educar em Revista*, Curitiba, Brasil, Edição Especial n. 1/2015, p. 17-33. Disponível em: https://www.scielo.br/j/er/a/JxHf3StKxjqmVTy97jbkccS/?format=pdf&lang=pt

FONSECA, Tina; CAMPIGLIA, Lucila. APREND.AI: Inteligência Artificial Generativa no ensino superior. *TECCOGS – Revista Digital de Tecnologias Cognitivas*, v. 28, p. 87–107. https://doi.org/10.23925/1984-3585.2023i28p87-107.

FREIRE, Paulo. *Conscientização*: teoria e prática da libertação: uma introdução ao pensamento de Paulo Freire. São Paulo: Cortez & Moraes, 1979.

GIANNAKOS, Michail, *et al*. The promise and challenges of generative AI in education. *Behaviour & Information Technology*, v. 43, n. 2, p. 10-22, 2024.

HOOKS, Bell. *Ensinando a transgredir*: a educação como prática de liberdade. São Paulo: WMF Martins Fontes, 2017.

Kaldaras L; Akaeze H.O.; Reckase M.D.Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, v. 9, p. 1-12, 2024.

LIMA, Cleosanice Barbosa; SERRANO, Agostinho. Inteligência artificial generativa e ChatGPT: investigação sobre impactos educacionais. *TransInformação*, v. 36, e236001, 2024.

LOIOLA, Alba; SACHETE, Andreia Santos; GOMES, Raquel Salcedo; GRANDI, Roges Horacio. IA generativa em competências discursivas na educação básica. *Revista Eletrônica de Educação*, v. 18, e6680021, 2024.

NOROOZI, Omid; SOLEIMANI, Saba; FARROKHNIA, Mohammadreza; BANIHASHEM, S. Kazem. Generative AI in education: pedagogical, theoretical and methodological considerations. *International Journal of Technology in Education*, v. 7, n. 2, p. 234-256, 2024.

SAÇÇO, Thays Alessandra Silva. *Influências do Ideb sobre a satisfação profissional na visão de professores do 1º ao 5º ano em seis escolas estaduais em Juiz de Fora / MG*. 2015. Tese (Doutorado em Educação) – Universidade Federal de Juiz de Fora, Faculdade de Educação, Juiz de Fora, 2015. Disponível em: https://repositorio.ufjf.br/jspui/handle/ufjf/3456.

SANTAELLA, Lucia. IA Generativa e o perfil semiótico-cognitivo do leitor iterativo. *Sociotramas*. 02 de janeiro de 2024. Disponível em: https://sociotramas.wordpress.com/2024/01/02/ia-generativa-e-o-perfil-semiotico-cognitivo-do-leitor-iterativo/

SANTOS, Antônio Bispo dos. *A terra dá, a terra quer*. São Paulo: Ubu Editora/PISEAGRAMA, 2023.

SILVA, Jackeline Sousa; SÁ, Cícera Alves Agostinho de. A Inteligência artificial generativa aplicada ao ensino inclusivo. *Revista Exitus*, v. 1, pág. 1-20, 2024.

TONBULOĞLU, B. State of research on e-assessment in education: A bibliometric analysis. *Turkish Online Journal of Distance Education*, v. 24, p. 276-301, 2023.

UNESCO. *Guia para a Inteligência Artificial Generativa na Educação e na Pesquisa*. Paris: UNESCO, 2024.

XIA, Qi, *et al*. A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, v. 21, p. 1-20, 2024.