

EnsinaAI: Geração de Questões com Qualidade Pedagógica Auditada por Catálogos Declarativos

Diego Grossmann^{1,2}, Thiago Reis Silva¹, Alana Oliveira², Mario Meireles Teixeira²

¹Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA)
São Luís – MA – Brasil

²Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brasil

{diego.grossmann, thiago.reis}@ifma.edu.br
{alana.oliveira, mario.meireles}@ufma.br

Abstract. *Large Language Models (LLMs) enable Automatic Item Generation at scale, but typically prioritize linguistic fluency over pedagogical soundness. We present a framework that formalizes pedagogical quality as 61 declarative rules across five catalogs (Evidence-Centered Design, Item-Writing Flaws (IWF), linguistic quality, technical correctness and rubric criteria) executed by a multi-agent generate–audit–regenerate loop with a deterministic decision engine. In an emergent study, 150 questions on Computer Networks generated by five distinct LLMs were audited by the linguistic (LIN) and IWF catalogs and reached 100% conformance to the engine’s criteria (approval being necessary, not sufficient, for pedagogical quality), with first-cycle approval ranging from 40% to 70% (χ^2 test, $p \approx 0.13$, no evidence of generator dependence at $\alpha=0.05$). The study also uncovered a systematic IWF–LIN asymmetry in the arbiter’s retention rate, suggestive of a GPT-4o-mini bias yet not separable from the higher intrinsic detectability of IWF flaws, persistent even in unrelated generators.*

Resumo. *Large Language Models (LLMs) viabilizam a Geração Automática de Itens em escala, mas tipicamente privilegiam fluência linguística em detrimento da fundamentação pedagógica. Apresentamos um framework que formaliza qualidade pedagógica como 61 regras declarativas em cinco catálogos (Evidence-Centered Design, Item-Writing Flaws — IWF, qualidade linguística, correção técnica e critérios de rubrica) executadas por um ciclo multiagente geração–auditoria–regeneração com motor de decisão determinístico. Em um estudo emergente, 150 questões de Redes de Computadores geradas por cinco LLMs distintos foram auditadas pelos catálogos linguístico (LIN) e IWF e atingiram 100% de conformidade com os critérios do motor (aprovação necessária, não suficiente, para a qualidade pedagógica), com aprovação em primeiro ciclo entre 40% e 70% (Qui-Quadrado, $p \approx 0,13$, sem evidência de dependência do modelo ao nível $\alpha = 0,05$). O estudo revelou ainda uma assimetria sistemática de retenção ($IWF \geq LIN$) do GPT-4o-mini como árbitro, sugestiva de viés, mas ainda não distinguível da maior detectabilidade intrínseca das falhas de Item-Writing Flaws — padrão persistente mesmo em geradores sem relação com ele.*

1. Introdução

A Geração Automática de Itens (*Automatic Item Generation*, AIG), historicamente abordada por métodos baseados em templates e regras de PLN [Kurdi et al. 2020, Silva Neo et al. 2024], ganhou destaque com *Large Language Models* (LLMs), que, contudo, tendem a privilegiar a fluência linguística em detrimento da fundamentação pedagógica [Tan et al. 2025], reproduzindo falhas clássicas da literatura de avaliação — como enunciados ambíguos associados à *construct-irrelevant variance* (variância alheia ao construto medido) [Haladyna and Downing 2004] — e introduzindo novas, como alucinações factuais [Huang et al. 2025]. O problema não está apenas nos modelos, mas na ausência de uma especificação formal do que constitui qualidade pedagógica: sem critérios explícitos e verificáveis, qualquer *pipeline* de geração opera no vácuo. Para o ensino de Redes de Computadores, revisões sistemáticas [Grosman et al. 2025, Aguiar et al. 2025] identificaram a escassez de IA aplicada à geração e avaliação de itens — nenhum dos dez trabalhos da Tabela 1 atua nesse domínio no ensino superior —, lacuna que motiva o desenvolvimento desse sistema.

Diante desse cenário, investiga-se se um processo estruturado e baseado em critérios explícitos pode mitigar os problemas identificados e revelar comportamentos dos modelos avaliativos. Formula-se, assim, a seguinte Questão de Pesquisa (QP): *Em que medida um ciclo de geração–auditoria–regeneração ancorado em catálogos declarativos de critérios pedagógicos (i) estabiliza a qualidade da saída entre geradores LLM heterogêneos e (ii) expõe vieses de juízes LLM individuais?*

Sendo assim, como contribuição, o *framework* converte cinco referenciais pedagógicos clássicos em catálogos declarativos de 61 regras, integrados a três *pipelines* multiagente — dos quais o fluxo de questão, auditado pelos catálogos LIN e IWF, é o avaliado empiricamente nesta rodada. A originalidade repousa em quatro pilares:

- **formalização declarativa** do conhecimento de avaliação (ECD, *Item-Writing Flaws*, qualidade linguística, correção técnica e critérios de rubrica) com IDs únicos, severidades e rastreabilidade à fonte primária;
- **motor de decisão determinístico** em Python puro — embora a detecção seja probabilística (LLMs), a decisão de fluxo (aprovar, regenerar ou finalizar com advertência) é tomada por código sobre a contagem de falhas por severidade, garantindo reprodutibilidade;
- **persistência granular** de cada falha em `AIAuditFlaw` com a tripla (`trace_id`, `flow_id`, `regeneration_round`), habilitando análise longitudinal e curadoria automática de *gold-standard* para ajuste fino supervisionado (*Supervised Fine-Tuning*, SFT) futuro;
- **evidência empírica de uma assimetria sistemática de retenção** ($IWF \geq LIN$) do GPT-4o-mini como árbitro, persistente em cinco geradores heterogêneos — nosso **achado emergente central** (Seção 6); um alerta à comunidade sobre o uso acrítico de LLMs como juízes de qualidade educacional.

O sistema é concebido como **triagem automatizada pré-revisão** que reduz a carga de revisão manual do docente ao filtrar itens malformados antes da curadoria humana, não como substituto do julgamento docente.

2. Fundamentação Teórica

O *Evidence-Centered Design* (ECD) [Almond et al. 2003, Mislevy et al. 2012] organiza a avaliação por meio de modelos conceituais interligados — com destaque para os modelos de Estudante, Evidência e Tarefa — utilizando ferramentas como algoritmos de avaliação ou rubricas para extrair observáveis e atualizar o modelo do estudante. Sua portabilidade é atestada por aplicações em domínios diversos, com aplicação direta ao *Cisco Networking Academy* [Behrens et al. 2010] para Redes de Computadores; em domínio distinto, Kubsch et al. [Kubsch et al. 2024] reforçam a validade psicométrica de detectores automatizados fundamentados em ECD. No contexto brasileiro, Jesus et al. [Jesus et al. 2020] alertam que avaliações educacionais em larga escala carecem de evidências de validade documentadas — reforço empírico à necessidade de critérios explícitos e verificáveis.

A taxonomia de diretrizes de construção de itens (*Item-Writing Guidelines*) de Haladyna et al. [Haladyna et al. 2002, Haladyna 2004] sistematiza décadas de consenso de especialistas. Análises com Teoria de Resposta ao Item (IRT) sobre mais de 7.000 questões STEM [Schmucker and Moore 2025] associam a presença de falhas estruturais a menor discriminação dos itens, com relação à dificuldade variável entre domínios; em um exame de ingresso a residência médica, aferido via Teoria Clássica dos Testes (CTT), apenas 5,8% dos itens estavam livres de falhas estruturais [Mohammed et al. 2025]. A detecção automática é viável: o *toolkit* SAQUET [Moore et al. 2024] atinge acurácia >94% na identificação de IWF. Moore et al. [Moore et al. 2023] mostraram que métodos baseados em regras superam o GPT-4 (91% vs. 79%), oferecendo interpretabilidade — evidência empírica que corrobora a escolha arquitetural por catálogos declarativos com motor determinístico. Quanto a critérios de rubrica, Popham [Popham 1997] distingue dimensões de qualidade transversais de critérios instrucionalmente fraudulentos (espelhamento da tarefa); Brookhart [Brookhart 2018] defende critérios gerais (que descrevem a qualidade de uma família de tarefas) em oposição a critérios específicos da tarefa, destacando ainda a necessidade de critérios substantivos focados na qualidade, enquanto a taxonomia revisada de Bloom [Anderson and Krathwohl 2001] ancora critérios em processos cognitivos. Em contexto brasileiro, Caporal et al. [Caporal et al. 2018] demonstraram que rubricas explícitas reduzem variabilidade entre corretores, e Carvalho et al. [Carvalho and Soares 2024] aplicam rubricas alinhadas à taxonomia de Bloom ao diagnóstico de aprendizagem em Computação.

Pesquisas com LLMs confirmam viabilidade da AIG [Doughty et al. 2024, Elkins et al. 2024], mas revisões apontam a priorização da fluência linguística em detrimento da fundamentação pedagógica [Tan et al. 2025]. O paradigma de auto-refinamento (Self-Refine [Madaan et al. 2023], *Reflexion* [Shinn et al. 2023], *Constitutional AI* [Bai et al. 2022]) e aplicações educacionais com auto-crítica iterativa (Edu-AgentQG [Jia et al. 2025], MCQG-SRefine [Yao et al. 2024]) demonstram ganhos consistentes do refinamento iterativo, e Kadavath et al. [Kadavath et al. 2022] inspiram a separação gerador-auditor ao evidenciar que a verificação escala mais rápido que a geração. Técnicas complementares de Engenharia de Prompt (*Chain-of-Thought*, CoT [Wei et al. 2022]; *Reasoning + Acting*, ReAct [Yao et al. 2023]) e Geração Aumentada por Recuperação (RAG) [Lewis et al. 2020] são instanciadas na arquitetura (Seção 5).

3. Trabalhos Relacionados

A partir da revisão, organizamos a literatura correlata em cinco blocos: (1) AIG psicométrica clássica [Gierl and Haladyna 2013], com templates fixos sem auditoria pós-geração; (2) geração com LLMs em domínios técnicos (Doughty [Doughty et al. 2024], Chan [Chan et al. 2025]), com métricas *ad hoc*; (3) sistemas agentivos e de auto-refinamento educacional (EduAgentQG [Jia et al. 2025], MCQG-SRefine [Yao et al. 2024], Multi-Examiner [Wang et al. 2025a, Wang et al. 2025b]), com auto-crítica iterativa; (4) avaliação automatizada (SAQUET [Moore et al. 2024, Moore et al. 2023], Gorgun e Bulut [Gorgun and Bulut 2025]), operando como *toolkits* externos sem fechar o ciclo; e (5) ECD com LLM (CoTAL [Cohn et al. 2025], Yang et al. [Yang et al. 2026]), focados em correção ou em Educação Básica. Nesta revisão, não encontramos trabalho que una um *catálogo declarativo formal* a um *motor de decisão determinístico* (Tabela 1): mesmo os sistemas mais próximos (e.g., Yang et al. [Yang et al. 2026] e Multi-Examiner [Wang et al. 2025b]) delegam a decisão de aprovar/regenerar ao julgamento de um LLM, o que limita a reprodutibilidade dessa decisão. A combinação *geração + LLM + ECD + multiagente + catálogo formal + auto-crítica + motor determinístico* aplicada a Redes no ensino superior permanece, assim, não coberta.

Tabela 1. Posicionamento do EnsinaAI frente a trabalhos relacionados.

Trabalho	Foco	LLM	ECD	Multi-ag.	Catál. formal	Self-crit.	Mot. det.
Gierl & Lai [Gierl and Haladyna 2013]	Ger.	–	–	–	T	–	–
Doughty et al. [Doughty et al. 2024]	Ger.	✓	–	–	–	–	–
Wang et al. [Wang et al. 2025a]	Ger.	✓	–	✓	–	✓	p
Multi-Examiner [Wang et al. 2025b]	Ger.	✓	–	✓	–	p	–
EduAgentQG [Jia et al. 2025]	Ger.	✓	–	✓	–	✓	p
MCQG-SRefine [Yao et al. 2024]	Ger.	✓	–	–	–	✓	p
Yang et al. [Yang et al. 2026]	Ger.	✓	✓	✓	–	✓	p
CoTAL [Cohn et al. 2025]	Corr.	✓	✓	–	–	–	–
Maus et al. [Maus et al. 2025]	Feed.	✓	✓	–	–	–	–
SAQUET [Moore et al. 2024]	Aval.	✓	–	–	p	–	–
EnsinaAI (este)	Ger.	✓	✓	✓	✓	✓	✓

Legenda: Ger. = geração; Corr. = correção; Feed. = feedback; Aval. = avaliação externa; T = baseado em templates. ✓ = componente central e explícito; p = parcial (presente de forma limitada, implícita ou em subcomponente); – = ausente. Para **Mot. det.**: ✓ = decisão de aprovar/regenerar tomada por código sobre limiares fixos; p = regra de parada/aceitação por limiar embutida no julgamento de um LLM. A tabela contrasta capacidades arquiteturais, não desempenho pedagógico validado.

4. O Framework Pedagógico-Computacional

A premissa central é que a qualidade pedagógica pode ser formalizada como restrições computacionalmente verificáveis - analogamente aos testes automatizados em software, uma questão não é considerada válida até que seja aprovada pelos auditores. A escolha tem suporte empírico: abordagens baseadas em regras superam LLMs na precisão de detecção de falhas formais [Moore et al. 2023, Moore et al. 2024]. Cada regra possui ID único, descrição do problema, orientações para auditor e gerador, *checklist* de verificação e rastreabilidade à fonte primária; a classificação como *adaptada*, *conceito similar* ou *própria* distingue derivações de princípios consolidados de contribuições arquiteturais inéditas.

Os cinco catálogos totalizam **61 regras** agnósticas à disciplina, com severidades (alta/A, média/M, baixa/B):

- **ECD** (19: 7A+9M+3B) — estrutura de observáveis (e.g., ECD-G3 Observável Subjetivo);
- **IWF** (14: 5A+6M+3B) — operacionalização de *item-writing flaws* (e.g., IWF-G1 Pista Gramatical, IWF-M5 Sobreposição de Alternativas);
- **LIN** (11: 4A+2M+5B) — problemas comunicativos não capturados pelo IWF (ambiguidade, comando confuso, lógica obscura);
- **TECH** (7: 3A+4M) — correção técnica, incluindo *Alucinação de Dados*, parametrizada pela disciplina;
- **CRIT** (10: 5A+4M+1B) — *design* de critérios de rubrica, incluindo *Espelhamento Algorítmico do Enunciado* (CRIT-H4 — *criterion-as-task*).

A troca de domínio exige apenas configuração de sujeito e tópico, sem alteração de regras.

Exemplo — IWF-G1 (Pista Gramatical, alta): “A concordância gramatical entre enunciado e alternativas não permite eliminar distratores?” (fonte: [Haladyna et al. 2002]). Catálogo completo em <https://doi.org/10.5281/zenodo.21310882>.

O ciclo geração–auditoria–regeneração tem propriedades conservadoras: (i) **terminação garantida** com limite N de ciclos e no máximo $N+1$ chamadas ao gerador; (ii) **aprovação parcial** quando restam apenas falhas médias após k ciclos; (iii) **contexto corretivo acumulativo**, com a lista de falhas do ciclo anterior vinculada aos IDs do catálogo; (iv) **self-critique com contestação**, em que o gerador emite um mapa $\{\text{ID-da-regra} \rightarrow \text{defesa}\}$ avaliado por auditores e árbitro, inspirado no paradigma *Constitutional AI* [Bai et al. 2022] sem delegar a decisão final ao gerador. A Figura 1 sintetiza o ciclo e os papéis envolvidos.

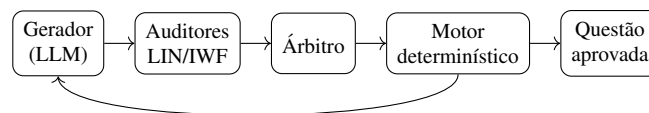


Figura 1. Ciclo geração–auditoria–regeneração.

5. Arquitetura e Implementação

5.1. Visão geral e fluxos

O EnsinaAI é uma plataforma *web* (*backend* Python/Django, *frontend* Next.js). O núcleo de geração usa *frameworks* modernos para orquestração multiagente (CrewAI) e abstração de provedores (LiteLLM), permitindo **auditoria cruzada entre modelos** configurável pelo `AIModelRegistry` — estratégia para mitigar o *self-enhancement bias* [Zheng et al. 2023]. O fornecimento de contexto baseia-se no paradigma de Geração Aumentada por Recuperação (RAG) [Lewis et al. 2020, Gao et al. 2023] com *embeddings* densos [Karpukhin et al. 2020, Reimers and Gurevych 2019], instanciado tecnicamente via PostgreSQL e pgvector. A plataforma expõe três fluxos de geração — Questão, Critérios de Rubrica e Observáveis ECD — com granularidades distintas de auditoria (Tabela 2). Os *prompts* são construídos como *templates* (Jinja2) e produzem saídas estruturadas (*structured output* via Pydantic). Para orquestrar o comportamento dos LLMs, os

templates instanciam uma *persona* pautada em princípios constitucionais [Bai et al. 2022], integram raciocínio em cadeia (CoT) [Wei et al. 2022] e ciclos de auto-crítica baseados no arcabouço *Reflexion* [Shinn et al. 2023], seguindo a sinergia entre raciocínio e ação (ReAct) [Yao et al. 2023]. O *design* também prevê mecanismos de defesa contra vulnerabilidades conhecidas como injeção de *prompt* [Greshake et al. 2023].

Tabela 2. Granularidade de cada auditor.

Auditor	Catálogo	Unidade de análise
Linguístico	LIN	Questão completa (enunciado + alternativas)
Psicométrico	IWF	Questão completa (enunciado + alternativas)
ECD	ECD	Um critério por vez (irmãos como contexto)
Técnico	TECH	Um observável por vez (dica de correção em foco)
Critérios	CRIT	Rubrica completa
Consolidador	—	Visão macro: rubrica + relatórios; cruzada

5.2. Motor de decisão determinístico e persistência

Os três fluxos compartilham a mesma lógica em Python puro: aprova quando $|alta| = |média| = 0$; aprova com advertência quando $|alta| = 0$ e $r \geq k$; regenera enquanto $r < N$; finaliza ao atingir $r=N$. Parâmetros: Questão ($N=3, k=2$), Critérios ($N=3, k=1$), Observáveis ECD ($N=2, k=1$), fixados pela saturação observada da convergência (ganho marginal desprezível além de N ciclos). A *detecção* de falhas é probabilística (LLMs); a *decisão de fluxo* é determinística (código).

A camada de monitoramento registra, em quatro entidades relacionais (AIAgentCall, AIAuditRun, AIAuditFlaw e AIToolExecution), todas as chamadas a LLMs, passes de auditor, falhas individuais (`flaw_type` $\rightarrow$ ID do catálogo, `severity`, sugestões) e chamadas RAG; logs de fluxo em `OperationalLog`. São persistidas *todas* as falhas, inclusive as descartadas pelo árbitro — habilitando análise de falsos positivos e curadoria de *gold-standard* para SFT. No conflito auditor–gerador, o árbitro avalia a defesa do gerador e retém ou descarta cada falha.

6. Avaliação Experimental

Protocolo. Foram geradas **150 questões** de Redes de Computadores (30 por gerador, 15 objetivas + 15 dissertativas, sobre sub-redes IPv4, OSPF, TCP/UDP, NAT e ARP) variando o modelo do gerador e mantendo os auditores fixos (árbitro: GPT-4o-mini; auditores LIN/IWF/ECD/TECH/CRIT: Llama-4-Maverick).

Modelos e parâmetros. Cinco geradores via OpenRouter (identificador exato com *snapshot*): **G1** DeepSeek V4 Pro (deepseek/deepseek-v4-pro-20260423), **G2** Qwen3-235B-A22B (qwen/qwen3-235b-a22b-04-28), **G3** Gemini 2.5 Flash (google/gemini-2.5-flash), **G4** Llama-4-Maverick (meta-llama/llama-4-maverick-17b-128e-instruct; colisão com auditores) e **G5** GPT-4o-mini (openai/gpt-4o-mini; colisão com árbitro). A geração usa temperatura 0,5 e `reasoning.effort=high` (esforço de raciocínio estendido); auditores operam a 0,2 (ECD a 0,0) e o árbitro a 0,1. Não se fixou *seed*: a geração é deliberadamente estocástica (o determinismo do sistema está no *motor de*

decisão, não na geração). Limite $N=3$ ciclos de regeneração por fluxo de questão. Métricas extraídas das tabelas `AIAGentCall`, `AIAGentRun` e `AIAGentFlaw`; dados completos em <https://doi.org/10.5281/zenodo.21310882>. Consideramos *válido* o flow cujo enunciado final foi persistido (30 por gerador); os demais (9/15/3/7/3) são re-execuções órfãs de jobs reiniciados, todos aprovados, sem esgotamento de N . Nesta rodada, atuaram os auditores LIN e IWF + árbitro nos fluxos de questão; TECH e CRIT operam nos fluxos de rubrica/observável ECD, não avaliados neste experimento.

6.1. Convergência: evidência preliminar de nivelamento

Todos os cinco geradores atingiram **100% de conformidade com os critérios do motor de decisão** (30/30 flows válidos cada, $|alta| = 0$ em todos): 139/150 não tiveram falhas médias remanescentes e 11/150 (G1 3, G3 1, G4 2, G5 5) foram *aprovações com advertência*, com apenas falhas de severidade média remanescentes, conforme a regra $r \geq k$ (Seção 5.2). A aprovação pelo motor é *necessária, não suficiente*, para a qualidade pedagógica. A aprovação em primeiro ciclo, porém, variou de 40% (G5) a 70% (G1) — Tabela 3. O teste de Qui-Quadrado sobre sucesso–retry ($\chi^2 \approx 7,16$, $gl=4$, $p \approx 0,13$) não encontrou evidência de dependência ao nível $\alpha = 0,05$; contudo, com $n=30$ por grupo, o poder estatístico é limitado para efeitos médios ($w \approx 0,2$) e a equivalência formal exigiria um TOST (*two one-sided tests*), não conduzido nesta rodada. Trata-se, portanto, de *evidência preliminar* de que o ciclo de auditoria estabiliza a *conformidade* final entre geradores heterogêneos — insuficiente para descartar heterogeneidade, dado o baixo poder.

Tabela 3. Convergência, retenção do árbitro, custo (fluxo de questão) e tempo por gerador ($n=30$ cada).

Métrica	G1	G2	G3	G4	G5
Aprovação 1º ciclo (%)	70,0	53,3	63,3	46,7	40,0
Ciclos médios	1,53	1,60	1,43	1,70	1,87
Retenção árbitro IWF (%)	89,5	94,7	100	95,0	90,9
Retenção árbitro LIN (%)	70,6	83,3	72,7	72,7	56,2
Custo médio/Q (US\$)	0,0158	0,0279	0,0204	0,0069	0,0080
Wall médio/Q (min)	3,07	5,22	1,19	0,88	2,40

6.2. Assimetria de retenção do árbitro: indício de viés

O árbitro **retém $\geq 89,5\%$ das falhas IWF** em todos os cinco geradores (Tabela 3), contra 56,2–83,3% das falhas LIN (taxa global 70,1%). Em números absolutos (reportadas→retidas): G1 17/19, G2 18/19, G3 12/12, G4 19/20, G5 30/33 — os volumes sustentam a tendência mesmo nos geradores com poucas falhas. O padrão persiste para G1 (DeepSeek), G2 (Qwen) e G3 (Gemini) — modelos sem qualquer relação com o árbitro GPT-4o-mini — o que *descarta self-enhancement bias como explicação*. Restam, em pé de igualdade, duas leituras concorrentes para a assimetria: (a) um **viés do GPT-4o-mini** que o leva a reter mais falhas na categoria IWF; ou (b) uma **maior detectabilidade intrínseca** das falhas IWF (estruturais) frente às LIN (estilísticas), que elevaria a retenção independentemente de qualquer viés. Os dados desta rodada não permitem distinguir as duas hipóteses — a discriminação exige variar o árbitro (rotação entre provedores, prevista como trabalho futuro). G5 (mesmo modelo do árbitro) registrou o maior volume de falhas IWF reportadas pelo auditor Llama-4-Maverick (33), o que ao menos reforça a recusa de um favorecimento intra-provedor.

6.3. Eficiência e modo *thinking*

O custo médio por questão varia $4\times$ entre G2 (US\$ 0,028) e G4 (US\$ 0,007), e o *wall* médio $6\times$ entre G2 (5,2 min) e G4 (0,9 min). Os testes de Kruskal-Wallis confirmam diferenças significativas em latência ($H = 77,14, p < 10^{-15}$) e custo ($H = 139,32, p < 10^{-28}$), após rejeição de normalidade via Shapiro-Wilk ($p < 10^{-67}$). O custo agregado do experimento (*pipeline* completo incluindo flows abandonados e auditoria) foi US\$ 9,95; considerando apenas os 150 flows válidos, US\$ 2,37 — demonstrando viabilidade econômica para produção institucional. Modelos com *thinking* ativo (G1 DeepSeek e G3 Gemini 2.5 Flash, ambos com `reasoning.effort=high`; razão completion/prompt (proxy de verbosidade) agregada sobre as chamadas de geração $\approx 0,3$) tiveram maior aprovação em 1º ciclo (70% e 63%) que os concisos (G4/G5: 40–47%; razão $\approx 0,035$), sugerindo *trade-off* entre verbosidade do gerador e convergência da auditoria.

6.4. Limitações

(i) TECH e CRIT operam nos fluxos ECD-observável e de rubrica, não diretamente no fluxo de questão — a auditoria factual completa de dissertativas depende do encadeamento dos três fluxos; (ii) a variabilidade de LLMs como juízes [Zheng et al. 2023] é mitigada pelo motor determinístico mas não eliminada nos auditores individuais; a geração não fixa *seed* (temperatura 0,5), sendo estocástica; (iii) a assimetria de retenção IWF do árbitro, para ser atribuída a viés, exige validação cruzada com revisores humanos, prevista como trabalho futuro; (iv) o domínio é um único contexto disciplinar — generalização carece de evidência empírica direta; (v) com $n=30$ por grupo, o poder estatístico é baixo para efeitos médios. A principal ameaça à validade, contudo, é a **circularidade**: a “aprovação” é definida e certificada pelo próprio motor, de modo que “100%” mede *conformidade* com os critérios, não qualidade pedagógica externamente validada — aprovação necessária, mas não suficiente — a validação por revisores humanos e a rotação do árbitro converteriam esses indícios em evidência.

7. Conclusão e Trabalhos Futuros

Este trabalho mostrou que qualidade pedagógica pode ser formalizada como restrições computacionalmente verificáveis e operacionalizada num ciclo geração–auditoria–regeneração como **triagem automatizada pré-revisão docente**. A execução de 150 questões com cinco geradores atingiu **100% de conformidade com os critérios do motor de decisão** (métrica autorreferente — validação por revisores humanos prevista como trabalho futuro), e o teste de Qui-Quadrado ($p \approx 0,13$) não encontrou evidência de dependência ao nível $\alpha = 0,05$. O achado emergente mais relevante é a assimetria de retenção do árbitro ($IWF \geq 89,5\%$ contra 56,2–83,3% em LIN, persistente em geradores sem relação com o GPT-4o-mini): *sugestiva* de viés, mas ainda não distinguível da maior detectabilidade intrínseca das falhas IWF — um alerta sobre o uso acrítico de LLMs como juízes de conteúdo educacional. Trabalhos futuros: (a) concordância (Kappa de Cohen) com juízes humanos ($\kappa > 0,6$); (b) rotação do árbitro via `AIModelRegistry`; (c) curadoria automática de *gold-standard* para SFT [Gorgun and Bulut 2025]; e (d) transferência para segundo domínio disciplinar.

Disponibilidade de Dados e Código

O catálogo consolidado das 61 regras (identificadores, severidades, orientações de detecção e geração, *checklist* e rastreabilidade) está disponível em <https://>

[//doi.org/10.5281/zenodo.21310882](https://doi.org/10.5281/zenodo.21310882); o código-fonte completo do *framework* em <https://github.com/diegogrosmann/EnsinaAI/tree/cbie2026-artigo-snapshot>.

Declaração de Uso de IA Generativa

Os autores utilizaram ferramentas de IA generativa como apoio à geração de código do sistema, ao planejamento do artigo, à geração de *scripts* de análise dos resultados e à revisão do texto. A **redação** do artigo não foi gerada por IA. A concepção, a metodologia, as análises e as conclusões são de responsabilidade exclusiva dos autores.

Referências

- Aguiar, R. K. B. d., Silva, T. R. d., Lima, R. V. d., Grosmann, D., Lima, B. V. A. d., and Oliveira, F. B. M. d. (2025). Métodos de ensino e aprendizagem em redes de computadores -uma revisão sistemática da literatura. *RENOTE*, 23(1):161–171.
- Almond, R., Mislevy, R., and Steinberg, L. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1):3–62.
- Anderson, L. W. and Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman, New York, complete edition edition.
- Bai, Y., Kadavath, S., Kundu, S., Askeel, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., and Levy, R. (2010). An evidence centered design for learning and assessment in the digital world. CRESST Report 778, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.
- Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, Volume 3 - 2018:22.
- Caporal, A., Prado, M. R., Bini, I., and Boller, C. (2018). Padronização da correção de questões dissertativas para professores de saúde coletiva do curso de medicina em uma instituição de ensino superior do oeste do paran . *Revista Meta: Avalia  o*, 10(28):54–74.
- Carvalho, F. S. and Soares, C. d. S. N. (2024). Avalia  o por rubricas seguindo a taxonomia de bloom no diagn stico de aprendizagem em algoritmos. In *Anais do Simp sio Brasileiro de Inform tica na Educa  o*.
- Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K., and Sze, G. K. (2025). Automatic item generation in various stem subjects using

- large language model prompting. *Computers and Education: Artificial Intelligence*, 8:100344.
- Cohn, C., Mohammed, N., Biswas, G., et al. (2025). Cotal: Human-in-the-loop prompt engineering for generalizable formative assessment scoring. *arXiv preprint arXiv:2504.02323*.
- Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., and Sakr, M. (2024). A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference, ACE '24*, pages 114–123, New York, NY, USA. Association for Computing Machinery.
- Elkins, S., Kochmar, E., Cheung, J. C. K., and Serban, I. (2024). How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23084–23091.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv e-prints*, page arXiv:2312.10997.
- Gierl, M. J. and Haladyna, T. M., editors (2013). *Automatic Item Generation: Theory and Practice*. Routledge, New York, NY.
- Gorgun, G. and Bulut, O. (2025). Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educational Measurement: Issues and Practice*, 44(1):96–107.
- Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, AISec '23*, pages 79–90, New York, NY, USA. Association for Computing Machinery.
- Grosmann, D., de Aguiar, R. K. B., da Silva, T. R., and Teixeira, M. M. (2025). Challenges, benefits, and perspectives on the use of virtual laboratories in computer networks teaching: A systematic mapping. In Sánchez, J., editor, *Nuevas Ideas en Informática Educativa*, volume 19, pages 87–97, Santiago de Chile.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 3rd edition.
- Haladyna, T. M. and Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1):17–27.
- Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3):309–333.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

- Jesus, G. R. d., Rêgo, R. M. d. L., and Souza, V. V. d. (2020). A questão da validade na avaliação educacional brasileira. *Ensaio: Avaliação e Políticas Públicas em Educação*.
- Jia, R., Zhang, M., Liu, F., Jiang, B., Kuang, K., and Dai, Z. (2025). Eduagentqg: A multi-agent workflow framework for personalized question generation. *arXiv preprint arXiv:2511.11635*.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. (2022). Language models (mostly) know what they know.
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *arXiv e-prints*, page arXiv:2004.04906.
- Kubsch, M., Grimm, A., Neumann, K., Drachsler, H., and Rummel, N. (2024). Using evidence-centered design to develop an automated system for tracking students physics learning in a digital learning environment. In Zhai, X. and Krajcik, J., editors, *Uses of Artificial Intelligence in STEM Education*, chapter 11, pages 230–249. Oxford University Press.
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9459–9474. Curran Associates, Inc.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 46534–46594.
- Maus, H., Tschisgale, P., Kieser, F., Petersen, S., and Wulff, P. (2025). Developing and evaluating a large language model-based automated feedback system grounded in evidence-centered design for supporting physics problem solving. *IEEE Transactions on Learning Technologies*.
- Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., and Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4(1):11–48.
- Mohammed, O. E. A. H., Srinivasamurthy, S. K., Bhat, R., Eltahir, Y. A. M. A., Elshamly, H. A. H., Mohammed, F., and Hamad, B. (2025). Postgraduate clinical residency: The impact of multiple-choice question quality on exam success rates. *Advances in Medical Education and Practice*, 16:1381–1397. PMID: 40786957.

- Moore, S., Costello, E., Nguyen, H. A., and Stamper, J. (2024). An automatic question usability evaluation toolkit. In Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., and Bittencourt, I. I., editors, *Artificial Intelligence in Education*, pages 31–46, Cham. Springer Nature Switzerland.
- Moore, S., Nguyen, H. A., Chen, T., and Stamper, J. (2023). Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In Viberg, O., Jivet, I., Muñoz-Merino, P. J., Perifanou, M., and Papathoma, T., editors, *Responsive and Sustainable Educational Futures*, volume 14200, pages 229–245, Cham. Springer Nature Switzerland.
- Popham, W. J. (1997). What’s wrong and what’s right-with rubrics. *Educational leadership*, 55(2):72–75.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Schmucker, R. and Moore, S. (2025). The impact of item-writing flaws on difficulty and discrimination in item response theory. *arXiv preprint arXiv:2503.10533*.
- Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv e-prints*, page arXiv:2303.11366.
- Silva Neo, A. V. B. d., Neo, G. d. S., Santos, M. D. F. d., Galvão Filho, K. J. A., and Freitas Junior, O. d. G. (2024). Revisão sobre a geração automática de questões na educação: Técnicas, conjuntos de dados e métricas de avaliação. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação (SBTI)*. SBTI.
- Tan, B., Armoush, N., Mazzullo, E., Bulut, O., and Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.
- Wang, J., Xiao, R., and Tseng, Y.-J. (2025a). Generating ai literacy mcqs: A multi-agent llm approach. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2, SIGCSETS 2025*, pages 1651–1652, New York, NY, USA. Association for Computing Machinery.
- Wang, Y., Yu, Z., Wang, Z., Yu, Z., and Wang, J. (2025b). Multi-examiner: A knowledge graph-driven system for generating comprehensive it questions with higher-order thinking. *Applied Sciences*, 15(10).
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24824–24837. Curran Associates, Inc.
- Yang, Y., Park, J., Zhou, Y., and Zhai, X. (2026). Developing a multi-agent system to generate next generation science assessments with evidence-centered design. *AI4STEM Education Center, University of Georgia*.

- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- Yao, Z., Parashar, A., Zhou, H., Jang, W. S., Ouyang, F., Yang, Z., and Yu, H. (2024). Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *arXiv preprint arXiv:2410.01553*.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 46595–46623. Curran Associates, Inc.