

Graph-Driven Exam Gen: Plataforma GraphRAG para Geração de Itens com Ancoragem na Matriz de Referência do ENEM

Jeová Leite¹, Arthur Lopes¹, Marcos Monteiro¹
Ig Ibert Bittencourt², Emanuel Marques Queiroga³, Aêda Monalliza Cunha de Sousa¹

¹Universidade de Pernambuco (UPE) — Campus Garanhuns
Rua Cap. Pedro Rodrigues, 105 — São José — 55.294-902 — Garanhuns — PE — Brazil

{jeova.leite, arthur.sampaio, marcos.pierre, aeda.sousa}@upe.br

²Universidade Federal de Alagoas (UFAL)

ig.ibert@ic.ufal.br

³Universidade Federal de Santa Maria (UFSM)

emanuelmqueiroga@gmail.com

Abstract. *LLM-based item generators produce grammatically correct questions but give no structural guarantee of alignment to ENEM's Reference Matrix [Tan et al. 2025]. We model the Matrix as a Neo4j property graph injected as structured curricular context (GraphRAG) and compare three generation conditions (graph context, few-shot without graph, no context) with GLM-5.1 at temperature 0. All 150 items (10 skills $\times$ 3 conditions $\times$ 5) reached structural conformance, and graph context yielded higher lexical coherence and specificity than both baselines. Pedagogical quality, however, still requires the expert evaluation set as next step.*

Resumo. *Sistemas de geração automática de itens baseados em LLMs produzem questões linguisticamente corretas, mas sem garantia de alinhamento à Matriz de Referência do ENEM [Tan et al. 2025]. A proposta modela a Matriz como grafo de propriedades Neo4j injetado como contexto curricular estruturado (GraphRAG) e compara três condições de geração (contexto do grafo, few-shot sem grafo, sem contexto) com GLM-5.1 a temperatura 0. Os 150 itens (10 habilidades $\times$ 3 $\times$ 5) apresentaram conformidade estrutural, e o contexto do grafo produziu maior coerência e especificidade lexical que as demais condições. A qualidade pedagógica, porém, ainda depende da avaliação por especialistas proposta como próxima etapa.*

1. Introdução

Com mais de quatro milhões de participantes anuais, o ENEM é o principal mecanismo de acesso ao ensino superior no Brasil. Sua validade pedagógica repousa na Matriz de Referência [INEP 2009]: cada questão avalia uma habilidade específica, dentro de uma

competência e área delimitadas — hierarquia que a Matriz formaliza em quatro áreas, subdivididas em competências e 120 habilidades numeradas. A modernização do exame passa pela construção de bancos de itens robustos para Teste Adaptativo Computadorizado (CAT), em que algoritmos de Teoria de Resposta ao Item (TRI) ajustam a dificuldade em tempo real [Jaloto and Primi 2024], exigindo rastreabilidade curricular item a item.

A demanda por materiais de preparação impulsionou ferramentas de geração automática de questões via LLMs, capazes de produzir itens linguisticamente corretos [Mucciaccia et al. 2025]. Contudo, revisão sistemática de 60 estudos documenta que muitos sistemas privilegiam qualidade linguística em detrimento da fundamentação educacional [Tan et al. 2025] — lacuna crítica dado que a Matriz impõe restrições hierárquicas explícitas entre competências, habilidades e objetos de conhecimento.

Trabalhos recentes integram grafos de conhecimento à geração de questões (KAQG [Chen and Shiu 2025], KNIGHT [Amanlou et al. 2026] e Wang et al. [Wang et al. 2026]), mas utilizam fontes de propósito geral (Wikipedia, Wikidata, corpora automáticos), sem ancoragem em matrizes curriculares normativas. A Matriz do ENEM, porém, não é um grafo de conhecimento de propósito geral: é um documento normativo emitido pelo INEP/MEC, com hierarquia fixada por política pública e escopo culturalmente específico ao Brasil.

A pergunta que motiva este trabalho é direta: *o contexto curricular estruturado em grafo de propriedades melhora a coerência pedagógica e a aderência ao formato ENEM de itens gerados por LLM em comparação à geração sem esse contexto?* Para investigá-la, modelamos a Matriz de Referência como grafo de propriedades Neo4j e construímos um *pipeline* GraphRAG que injeta esse contexto no *prompt*. A hipótese é que a ancoragem explícita na Matriz produz itens de maior qualidade pedagógica, confirmável por avaliação de especialistas.

Esta proposta se distingue das existentes por ancorar a geração na Matriz normativa oficial, não em fontes de propósito geral. O delineamento quasi-experimental mantém modelo e template fixos entre as condições, garantindo replicabilidade. O protótipo conta com 50 testes automatizados e métricas auditáveis; a avaliação pedagógica por especialistas, com protocolo na Seção 6, é a próxima etapa.

2. Referencial Teórico

2.1. O ENEM e sua Matriz de Referência

A rastreabilidade habilidade–questão é requisito pedagógico intrínseco da Matriz [INEP 2009]: cada item deve avaliar uma habilidade específica dentro da hierarquia área–competência. No paradigma CAT, a TRI seleciona dinamicamente o próximo item ajustando a dificuldade em tempo real, o que exige bancos com centenas de itens de parâmetros psicométricos conhecidos e alinhamento curricular verificável [Jaloto and Primi 2024] — tornando a geração automatizada com ancoragem na Matriz uma necessidade prática.

2.2. Geração Automática de Itens

Com a difusão dos LLMs (*Large Language Models*) pré-treinados em corpora massivos [Brown et al. 2020], a produção automatizada de questões de múltipla escolha ganhou

novo impulso na Inteligência Artificial Educacional: *prompts* com exemplos no próprio contexto (*few-shot*) viabilizam a geração de itens sem ajuste fino do modelo. A geração, contudo, é intrinsecamente probabilística — saídas variam entre execuções —, propriedade que exige controle explícito de temperatura para garantir replicabilidade experimental (Seção 4.3).

Kurdi et al. [Kurdi et al. 2020] revisam sistematicamente 93 estudos sobre geração de questões, identificando abordagens baseadas em regras, *templates*, PLN e modelos neurais. Tan et al. [Tan et al. 2025] documentam, em revisão de 60 estudos com LLMs, que a maioria privilegia a qualidade linguística em detrimento da fundamentação educacional: itens gramaticalmente corretos, sem mecanismo estrutural que garanta alinhamento à habilidade curricular-alvo. No contexto brasileiro, Mucciaccia et al. [Mucciaccia et al. 2025] mostram que MCQs geradas por LLM requerem revisão especializada; Quincozes et al. [Quincozes et al. 2025] evidenciam que a qualidade depende da estratégia de engenharia de *prompt*; Santos et al. [Santos et al. 2025] constataam limitações conceituais e semânticas mesmo em modelos otimizados — reforçando, em conjunto, a necessidade de validação humana especializada.

2.3. Grafos de Conhecimento, Neo4j e LLMs na Geração de Questões

Grafos de propriedades estruturam domínios como nós e arestas com atributos diretos, sem esquema fixo, e permitem travessia de caminhos arbitrários em tempo constante por salto [Robinson et al. 2015] — ideais para estruturas curriculares de profundidade variável. O Neo4j armazena dados nativamente como grafo, com integridade transacional e integração ao ecossistema GraphRAG, o que motivou sua adoção: a Matriz de Referência do ENEM, hierárquica, normativa e com metadados de auditoria, mapeia-se diretamente ao modelo de propriedades (áreas, competências, habilidades e objetos de conhecimento como nós; relações curriculares como arestas tipadas). A integração com LLMs também reduz alucinações ao fornecer ancoragem factual verificável [Agrawal et al. 2024, Lavrinovics et al. 2025], mecanismo crítico em GAI (Geração Automática de Itens), onde erros factuais comprometem a validade da avaliação.

Desde abordagens baseadas em *templates* [Seyler et al. 2017], a geração de questões a partir de grafos evoluiu ao paradigma RAG (*Retrieval-Augmented Generation*), que enriquece o *prompt* com fragmentos recuperados de uma base externa. O GraphRAG [Han et al. 2025, Li et al. 2025] vai além, substituindo a busca vetorial por travessia de grafo e habilitando raciocínio multi-hop sobre relações estruturadas — vantagem direta para hierarquias curriculares. KAQG [Chen and Shiu 2025], KNIGHT [Amanlou et al. 2026] e Wang et al. [Wang et al. 2026] aplicam essa lógica à geração de questões, mas operam sobre fontes de propósito geral; nenhum ancora a geração em uma matriz curricular normativa oficial.

Trabalhos de conferências de Informática na Educação reforçam esse padrão. Scaria et al. [Scaria et al. 2024] (AIED 2024) geram questões em níveis da Taxonomia de Bloom, mas o alinhamento vem de instrução no *prompt*, sem representação estrutural persistente. Lai et al. [Lai et al. 2026] (LAK 2026) e Wei et al. [Wei et al. 2025] (EDM 2025) constroem estruturas curriculares *bottom-up* a partir do desempenho de estudantes, não de uma matriz normativa preexistente. Liang et al. [Liang et al. 2026] propõem o K12-KGraph, grafo de livros didáticos voltado a ajuste fino, não à geração com restrição

estrutural em tempo real. Lu e Wang [Lu and Wang 2024] (L@S 2024) simulam perfis de estudantes para apoiar a avaliação de itens, alternativa complementar à avaliação por especialistas adotada aqui (Seção 6). Nenhum ancora a geração em uma matriz curricular oficial, pública e normativa como a do ENEM.

3. Arquitetura Implementada

3.1. Visão Geral

A plataforma é composta por três camadas (Figura 1a): interface de usuário, camada de serviço (API REST) e camada de dados centrada no Neo4j. Toda geração passa pela camada de dados: o grafo impõe coerência curricular de forma estrutural, não como pós-processamento.

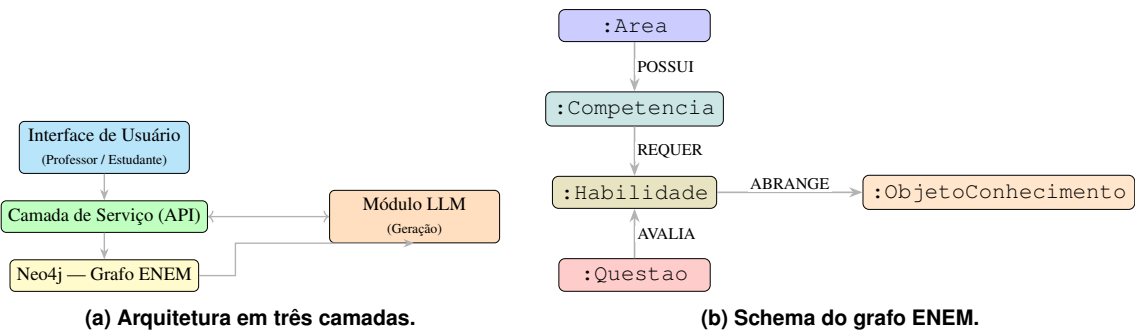


Figura 1. Arquitetura da plataforma (esq.) e schema do grafo de propriedades ENEM (dir.).

Fonte: elaborado pelos autores.

Cada nó `:Questao` armazena enunciado, alternativas, gabarito, ano de origem, condição de geração (ENEM, `com_grafo`, `com_fewshot_sem_grafo` ou `sem_grafo`) e a flag `validacao_estrutural_ok`. A aresta `AVALIA` conecta cada questão a uma ou mais habilidades (Figura 1b), garantindo rastreabilidade plena entre item e justificativa pedagógica — relação que distingue a abordagem de repositórios indexados apenas por tema ou disciplina.

3.2. População do Banco de Dados

O grafo é carregado em duas fases por instruções Cypher `MERGE` idempotentes. A primeira carrega a estrutura normativa completa: nós `:Area`, `:Competencia` e `:Habilidade` para as quatro áreas e 120 habilidades da Matriz [INEP 2009], extraídos *verbatim* do documento oficial, com metadados de auditoria; os nós `:ObjetoConhecimento` foram curados pelos autores a partir da descrição oficial de cada habilidade. O recorte experimental — geração e avaliação controladas — permanece restrito às habilidades H17–H26 (Ciências da Natureza, competências C5–C7), escolha deliberada para controle da variável experimental (Seção 4). A segunda fase ingere 35 questões reais do ENEM 2023 como exemplos *few-shot*, criando o nó `:Questao` com aresta `AVALIA` (procedência e critérios de mapeamento na Seção 4.5).

3.3. Pipeline de Geração de Avaliações

O *pipeline* recebe a habilidade-alvo e a condição experimental, consulta o grafo via Cypher percorrendo `Area→Competencia→Habilidade←Questao` para construir o `context_block` seguindo GraphRAG [Han et al. 2025, Li et al. 2025], monta

o `PROMPT_BASE` (Seção 4.4), invoca o LLM com os parâmetros da Tabela 1 e processa a resposta em JSON. Após validação estrutural (Seção 4.5), o item é persistido como nó `:Questao com condicao_geracao e validacao_estrutural_ok`.

4. Metodologia

4.1. Questão de Pesquisa e Hipótese Central

A questão de pesquisa e a hipótese central foram enunciadas na Seção 1; esta seção as operacionaliza no desenho experimental a seguir.

4.2. Desenho Experimental

O experimento adota delineamento quasi-experimental (sem atribuição aleatória, com controle por equivalência de protocolo) com três condições paralelas. A **variável independente** é a condição de geração: *com_grafo* (hierarquia curricular completa mais exemplos *few-shot*), *com_fewshot_sem_grafo* (apenas exemplos, sem objetos de conhecimento) e *sem_grafo* (somente código e descrição da habilidade) — a condição intermediária isola o efeito dos exemplos do efeito da ancoragem curricular. A **variável dependente** é a qualidade pedagógica dos itens, avaliada em cinco dimensões via escala Likert 1–5 (Seção 6). Para cada habilidade do recorte H17–H26, foram gerados cinco itens por condição: 150 no total ($10 \times 3 \times 5$).

4.3. Especificação do LLM e Controle de Estocasticidade

A Tabela 1 apresenta os parâmetros de inferência. A temperatura 0 adota decodificação gulosa (*greedy decoding*): o token de maior probabilidade é selecionado deterministicamente a cada passo, de modo que, dado o mesmo modelo, versão de API e *prompt*, a saída é idêntica entre execuções — condição necessária para comparação controlada, já que temperatura > 0 inviabilizaria replicação.

Tabela 1. Parâmetros de inferência do LLM (todos os não listados em valor padrão da API). Fonte: elaborado pelos autores.

Parâmetro	Valor	Parâmetro	Valor
Provedor/Modelo	GLM-5.1 (Zhipu AI)	Temperatura	0
ID da API	glm-5.1	max_tokens	8.192
Tentativas máx.	2	Data de acesso	Jul. 2026

4.4. Template de *Prompt* e Garantia de Equivalência

O sistema utiliza um único template `PROMPT_BASE` com quatro blocos fixos; apenas o `{context_block}` varia entre condições:

```
[SISTEMA] Você é especialista em elaboração de questões ENEM.  
Gere um item de múltipla escolha alinhado à habilidade informada.  
Responda exclusivamente em JSON, sem texto adicional.  
[CONTEXTO] {context_block}  
[INSTRUÇÕES] Enunciado situacional (não pode ser definição direta).  
5 alternativas A-E. Gabarito único. Sem LaTeX. Linguagem Ensino Médio.  
Justificativa com vínculo explícito à habilidade-alvo.  
[SAÍDA] {"enunciado": "...", "alternativas": {"A": "...", "B": "...",  
"C": "...", "D": "...", "E": "..."}, "gabarito": "X", "justificativa": "...}
```

O `context_block` diferencia as condições: *com_grafo* injeta a hierarquia completa (área, competência, habilidade e objetos de conhecimento) mais exemplos; *com_fewshot_sem_grafo* injeta só os exemplos; *sem_grafo* injeta só código e descrição da habilidade. Testes automatizados verificam que a única diferença entre *prompts* de condições distintas é esse bloco.

4.5. Validação Estrutural e Base de Itens

Todo item gerado passa por cinco critérios verificados programaticamente: (i) exatamente cinco alternativas A–E; (ii) gabarito em {A,B,C,D,E}; (iii) sem alternativas duplicadas; (iv) enunciado com ≥ 50 caracteres; (v) justificativa com ≥ 100 caracteres. Itens reprovados recebem `validacao_estrutural_ok=False`, são preservados no grafo para análise e excluídos do *export* para avaliação humana.

O repositório foi populado com 35 questões reais do ENEM 2023 (Ciências da Natureza, H17–H26), transcritas dos cadernos de prova oficiais (aplicação regular e re-aplicação PPL) e conferidas contra o gabarito oficial do INEP. O mapeamento à habilidade é por análise de conteúdo dos autores, não extração automática — os microdados públicos do INEP não trazem esse campo no nível de item individual —, e cobre de forma desigual o recorte (1 a 6 itens por habilidade, conforme disponibilidade de material verificável). Essas questões servem como exemplos *few-shot* em *com_grafo* e *com_fewshot_sem_grafo* e não integram o corpus de avaliação pedagógica. A coerência lexical de um item i com habilidade h é $C(i, h) = |T(i) \cap O(h)|/|T(i)|$, com $T(i)$ os *tokens* do item (enunciado e alternativas) e $O(h)$ os de seus objetos de conhecimento — ambos tokenizados em minúsculas, sem pontuação, com lista curta de *stopwords* do português removida. A especificidade $E(i, h) = C(i, h_{alvo})/\bar{C}(i, h_j)_{j \neq alvo}$ mede o quanto o item discrimina sua habilidade-alvo das demais, comparando apenas contra as demais habilidades do recorte experimental.

5. Resultados Emergentes

5.1. Implementação e Cobertura de Testes

O protótipo foi implementado em Python com API REST e Neo4j 5. Os 50 testes automatizados (43 unitários com *mocks* e 7 de integração contra o grafo populado) verificam os critérios de validação estrutural (Seção 4.5), a equivalência dos *prompts* entre condições (Seção 4.4), a idempotência das operações Cypher no grafo e o fluxo *end-to-end*. O ambiente é reproduzível com dependências declaradas e *prompts* versionados em repositório.

5.2. Validação Operacional do Pipeline

O *pipeline* gerou 150 itens (10 habilidades $\times$ 3 condições $\times$ 5 itens), todos com 100% de conformidade estrutural após reexecução idempotente do script (uma tentativa inicial falhou por resposta vazia da API; a célula foi preenchida sem edição manual de dados, graças ao MERGE). *Tokens* de entrada/saída e latência médios crescem de *sem_grafo* (471/2.986/48,7 s) para *com_fewshot_sem_grafo* (699/3.029/49,8 s) e *com_grafo* (814/3.151/51,6 s), refletindo o contexto curricular adicional injetado no *prompt*. A Tabela 2 apresenta os indicadores de coerência lexical e especificidade curricular (Seção 4.5).

Tabela 2. Coerência lexical e especificidade curricular por condição, média $\pm$ desvio-padrão sobre os 50 itens da condição ($n=5$ itens/habilidade $\times$ 10 habilidades). Mediana da especificidade incluída à parte por ser uma razão sensível a valores extremos (Seção 5). Fonte: elaborado pelos autores.

Condição	Coerência	Especificidade	Espec. (mediana)
com_grafo	0,047 $\pm$ 0,032	8,22 $\pm$ 7,66	5,70
com_fewshot_sem_grafo	0,030 $\pm$ 0,023	5,27 $\pm$ 4,71	4,50
sem_grafo	0,032 $\pm$ 0,021	5,21 $\pm$ 5,18	4,50

Em agregado, *com_grafo* supera as demais condições em ambas as métricas, consistente com a hipótese central — mas com dispersão alta: o desvio-padrão da especificidade é próximo ou maior que a própria média (Tabela 2), pois $E(i, h)$ é uma razão sensível a denominadores pequenos, e a mediana (bem abaixo da média nas três condições) é a leitura mais robusta desse indicador. O padrão por habilidade também é mais nuançado que uma ordenação uniforme: *com_grafo* supera *sem_grafo* em coerência em 8 das 10 habilidades, com ganhos maiores em H20 (+138%) e H26 (+105%); a exceção mais expressiva é H18 ($\sim 70\%$ a favor de *sem_grafo*, ainda pequena em termos absolutos), com H21 apenas marginal ($\sim 10\%$). Notavelmente, *com_fewshot_sem_grafo* — exemplos sem a lista de objetos de conhecimento — não superou *sem_grafo* em nenhuma métrica agregada, sugerindo que exemplos isolados não deslocam de forma confiável o vocabulário em direção aos objetos da habilidade-alvo; o ganho de *com_grafo* parece depender da combinação de hierarquia explícita e exemplos. Como análise exploratória complementar — possível ao agrupar os 50 itens/condição entre as 10 habilidades, o que introduz dependência intra-habilidade e deve ser lido com cautela —, a diferença *com_grafo* vs. *sem_grafo* corresponde a efeito médio (Cohen's $d = 0,55$; IC 95% [0,004; 0,025] para a diferença de médias), contra efeito nulo ($d = -0,08$) para *com_fewshot_sem_grafo* vs. *sem_grafo*. Dado $n=5$ por célula, os valores constituem tendência descritiva preliminar; a avaliação por especialistas (Seção 6) determinará se a diferença lexical observada se traduz em qualidade didática.

5.3. Ilustração Qualitativa Comparativa

Inspeção qualitativa dos 150 itens revela que a instrução "Sem LaTeX" do PROMPT_BASE foi seguida em 100% dos itens nas três condições, ao contrário do que uma versão preliminar deste trabalho hipotetizara — o efeito do grafo aparece, em vez disso, na *seleção do subtema* dentro da habilidade. Para H20 (maior ganho de coerência), o item *com_grafo* ancorou-se em órbita e microgravidade da ISS, alinhado aos objetos "gravitação universal" e "movimentos de corpos celestes"; o item *com_fewshot_sem_grafo* derivou para difusão de corante em água parada, tema tangente mas fora dos objetos curados; o *sem_grafo* também convergiu para mecânica celeste (órbita de cometas), sugerindo que o pré-treinamento já favorece esse subtema independentemente do contexto. Já em H21, onde *sem_grafo* venceu na coerência média, as três condições convergiram ao mesmo exemplo canônico (fogão por indução) com redação quase intercambiável — evidência de que, para habilidades com exemplar fortemente representado no pré-treinamento, o grafo tem pouco espaço para diferenciar a saída. Os valores absolutos de coerência (0,01–0,11) mesmo em itens tematicamente adequados também expõem limitação do indicador: o LLM parafraseia os objetos de conhecimento em vez de reutilizar seu

vocabulário literal, de modo que baixa sobreposição lexical não implica baixa adequação de conteúdo. Fragilidades de *design* didático (distratores facilmente elimináveis) persistiram em todas as condições. Esses achados são qualitativos e dependem de confirmação por avaliação sistemática.

6. Plano de Validação Empírica

A viabilidade técnica do *pipeline* está demonstrada; o que permanece em aberto é a hipótese central: o grafo de fato melhora a qualidade pedagógica? Responder a essa pergunta exige juízes especialistas em elaboração de itens ENEM avaliando as três condições de forma cega.

6.1. Protocolo de Avaliação por Juízes Especialistas

A avaliação seguirá protocolo *human-in-the-loop* [Kim et al. 2025, Runge et al. 2024] com dois a três juízes especialistas em elaboração de itens ENEM. A rubrica avaliará em escala Likert 1–5 cinco dimensões: alinhamento à habilidade-alvo, adequação dos objetos de conhecimento, plausibilidade dos distratores, adequação ao Ensino Médio e aderência ao formato ENEM; com registro adicional de erros factuais em escala binária. A anonimização entre condições já está implementada no *pipeline*; o protocolo é cego para os juízes.

6.2. Análise Estatística

A concordância inter-juízes será calculada por Kappa de Cohen (classificações binárias) ou Coeficiente de Correlação Intraclass (Likert). A comparação entre as três condições usará Kruskal-Wallis, com comparações *post-hoc* por Mann-Whitney U e correção de Bonferroni para múltiplos testes, apropriado para amostras pequenas e dados ordinais; complementado por análise qualitativa dos comentários dos juízes.

7. Considerações Finais

A Matriz de Referência do ENEM, enquanto documento normativo com hierarquia explícita área–competência–habilidade, oferece uma base formal rara em sistemas de GAI: a coerência curricular pode ser verificada por consulta ao grafo, não inferida do texto gerado — tornando o alinhamento pedagógico rastreável via aresta AVALIA, ao contrário de abordagens sobre fontes de propósito geral.

A inspeção qualitativa reforça esse quadro: o efeito do grafo concentra-se na *seleção do subtema* dentro da habilidade-alvo (Seção 5), não na conformidade de formato, e é mais evidente onde o pré-treinamento do modelo não converge naturalmente para um exemplar único. Se essas diferenças lexicais correspondem a maior qualidade pedagógica é o que a avaliação por especialistas (Seção 6) determinará.

Este trabalho tem limitações que condicionam a leitura de todos os resultados acima, não apenas a fecham. Primeiro: a hipótese central — de que o grafo melhora a *qualidade pedagógica* — permanece não testada por especialistas; os resultados são inteiramente automatizados e caracterizam *viabilidade técnica*, não qualidade didática, validade de conteúdo ou adequação psicométrica. Ler a Tabela 2 como demonstração de superioridade pedagógica extrapola o que os dados sustentam. Segundo, $n = 5$ por célula limita os resultados a uma tendência descritiva, sem poder de inferência estatística.

Terceiro, o recorte em Ciências da Natureza (H17–H26) é deliberado para controle experimental; generalizar para as demais áreas — já presentes na estrutura completa do grafo, mas fora do escopo deste experimento — é passo seguinte. Quarto, a escolha do GLM-5.1 (Zhipu AI), por custo e desempenho, exclui comparação com outros modelos, prevista em fase posterior. Bancos de itens calibrados são pré-requisito para o ENEM adaptativo [Jaloto and Primi 2024]; a rastreabilidade curricular aqui implementada é condição que grafos de propósito geral não satisfazem, mas não deve ser confundida com a validação pedagógica ainda pendente.

Declaração sobre uso de Inteligência Artificial

O modelo de linguagem GLM-5.1 (Zhipu AI) constitui objeto de estudo deste trabalho e foi empregado exclusivamente na geração dos itens avaliados, conforme a Seção 4. No que se refere à redação deste artigo: ferramentas de IA generativa foram utilizadas como apoio à revisão linguística e à clareza do texto; a concepção, a análise e as conclusões são de responsabilidade integral dos autores, que revisaram e validaram todo o conteúdo.

Disponibilidade de Artefatos

O código-fonte do protótipo (API REST e *pipeline* de geração), o *schema* do grafo, os *prompts* versionados e a suíte de testes automatizados estão disponíveis em <https://github.com/ArthurSampaio13/graph-gen>. As 35 questões reais do ENEM 2023 usadas como exemplos *few-shot* não são redistribuídas por serem de titularidade do INEP; as instruções de reprodução com os dados oficiais estão documentadas no README.

Agradecimentos

Este trabalho foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio do processo nº 445016/2024-8; pelo projeto Ia.Edu – Instituto Nacional de Inteligência Artificial em Educação Desplugada (processo nº 4084883/2024-5); e pela Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL), por meio do processo nº FAPEAL60030.0000001481/2025.

Referências

- Agrawal, G., Kumarage, T., Alghamdi, Z., and Liu, H. (2024). Can knowledge graphs reduce hallucinations in LLMs? A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3947–3960. Association for Computational Linguistics.
- Amanlou, M., Moghaddam, E. S., Jafari, Y. A., Noori, M., Farsi, F., and Bahrak, B. (2026). KNIGHT: Knowledge graph-driven multiple-choice question generation with adaptive hardness calibration. In *Proceedings of the Third Conference on Parsimony and Learning (CPAL 2026)*. arXiv:2602.20135.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Chen, C. H. and Shiu, M. F. (2025). KAQG: A knowledge-graph-enhanced RAG for difficulty-controlled question generation. *IEEE Access*. Early access.

- Han, H., Ma, L., Wang, Y., Shomer, H., Lei, Y., Qi, Z., Guo, K., Hua, Z., Long, B., Liu, H., Aggarwal, C. C., and Tang, J. (2025). RAG vs. GraphRAG: A systematic evaluation and key insights. *arXiv preprint*, arXiv:2502.11371.
- INEP (2009). Matriz de referência para o ENEM 2009. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP/MEC), Brasília.
- Jaloto, A. and Primi, R. (2024). Enem de próxima geração com menos itens e alta confiabilidade usando CAT. *Estudos em Avaliação Educacional*, 35:e10142.
- Kim, E., Li, S., Khalil, S., and Shin, H. J. (2025). STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 920–930. Association for Computational Linguistics.
- 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.
- Lai, J. W., Ho, S. Y., and Lim, F. S. (2026). Learning analytics for assessment preparation: Constructing graphs to guide higher-order question generation. In *Proceedings of the 16th International Learning Analytics and Knowledge Conference (LAK 2026)*.
- Lavrinovics, E., Biswas, R., Bjerva, J., and Hose, K. (2025). Knowledge graphs, large language models, and hallucinations: An NLP perspective. *Web Semantics: Science, Services and Agents on the World Wide Web*, 85:100844.
- Li, M., Miao, S., and Li, P. (2025). Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*. arXiv:2410.20724.
- Liang, H., Lin, Q., Han, Z., Ma, X., Wong, Z. H., Qiang, M., Sun, L., and Zhang, W. (2026). K12-KGraph: A curriculum-aligned knowledge graph for benchmarking and training educational LLMs. *arXiv preprint*, arXiv:2605.09635.
- Lu, X. and Wang, X. (2024). Generative students: Using LLM-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S 2024)*. ACM.
- Mucciaccia, S. S., Paixão, T. M., Mutz, F., Badue, C. S., De Souza, A. F., and Oliveira-Santos, T. (2025). Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260.
- Quincozes, C. B., Molinos, D., Araújo, R. D., Quincozes, S., and Guedes, G. T. A. (2025). Engenharia de prompt para a geração automatizada de questões assistida por LLMs: Uma análise comparativa. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*. SBC.
- Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data*. O’Reilly Media, Sebastopol, CA, 2nd edition.

- Runge, A., Attali, Y., LaFlair, G. T., Park, Y., and Church, J. (2024). A generative AI-driven interactive listening assessment task. *Frontiers in Artificial Intelligence*, 7.
- Santos, M. M., Barros, A. P., Santos, E., Silva, J. G. d., Isotani, S., Bittencourt, I. I., Marcario, V., Rodrigues, L., and Dermeval, D. (2025). Geração de questões com LLMs leves: Um estudo inicial sobre a percepção de educadores. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 1635–1646. SBC.
- Scaria, N., Chenna, S. D., and Subramani, D. N. (2024). Automated educational question generation at different bloom’s skill levels using large language models: Strategies and evaluation. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED 2024)*, Lecture Notes in Computer Science, pages 165–179. Springer.
- Seyler, D., Yahya, M., and Berberich, K. (2017). Knowledge questions from knowledge graphs. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017)*, pages 11–18, Amsterdam, The Netherlands.
- Tan, B., Armoush, N., Mazzullo, E., Bulut, O., and Gierl, M. J. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.
- Wang, Y., Wei, T., Li, Q., and Zeng, L. (2026). Beyond static question banks: Dynamic knowledge expansion via LLM-automated graph construction and adaptive generation. *Proceedings of the VLDB Endowment*, 14(1). arXiv:2602.00020.
- Wei, Y., Carvalho, P., and Stamper, J. (2025). KCluster: An LLM-based clustering approach to knowledge component discovery. In *Proceedings of the 18th International Conference on Educational Data Mining (EDM 2025)*. International Educational Data Mining Society. arXiv:2505.06469.