

Simulação de Proficiência Discente do ENEM via Grandes Modelos de Linguagem: Uma Análise da Aderência à TRI

Caio César Farias da Silva¹, Roberta Macedo Marques Gouveia¹
Maria da Conceição Moraes Batista¹, Gabriel Alves¹

¹Universidade Federal Rural de Pernambuco (UFRPE)
Recife - PE - Brasil

{caio.fsilva, roberta.gouveia, maria.cmbatista, gabriel.alves}@ufrpe.br¹

Abstract. *Item calibration in Item Response Theory is an operationally complex and costly process, requiring a significant number of respondents. This paper proposes the use of Large Language Models to simulate responses to ENEM exam questions, combining prompt engineering and data serialization formats to evaluate both psychometric adherence and computational cost. The results indicate the viability of the approach, preserving the difficulty ordering of the items. The TOON format reduced token consumption by approximately 25% without significant loss of quality. The analysis demonstrated that gains in psychometric fidelity and operational efficiency can be mutually achieved.*

Resumo. *A calibração de itens na Teoria da Resposta ao Item é um processo operacionalmente complexo e oneroso, exigindo uma quantidade significativa de respondentes. Este trabalho propõe o uso de Grandes Modelos de Linguagem para a simulação de respostas a questões do ENEM, combinando engenharia de prompt e formatos de serialização de dados, avaliando a aderência psicométrica e o custo computacional. Os resultados indicam a viabilidade da abordagem, com a preservação da ordenação de dificuldade dos itens. O formato TOON reduziu o consumo de tokens em aproximadamente 25% sem perda expressiva de qualidade. A análise demonstrou que ganhos de fidelidade psicométrica e eficiência operacional podem ser obtidos de forma mútua.*

1. Introdução

A manutenção da qualidade e equidade em sistemas de avaliação em larga escala pressupõe rigor metodológico na construção e calibração de itens avaliativos, visto que a integridade depende da sua capacidade de mensurar a proficiência do avaliado sem vieses que comprometam a confiabilidade dos resultados [Pasquali 2003]. Dada a função estratégica do Exame Nacional do Ensino Médio (ENEM) no acesso ao ensino superior e o expressivo contingente anual de candidatos, torna-se imperativo assegurar a clareza dos critérios e do conteúdo programático [Sobrinho and Cavalcanti 2023].

Para garantir integridade estatística necessária a exames dessa magnitude e aos bancos de itens, utiliza-se a Teoria da Resposta ao Item (TRI), modelo matemático que permite estimar a proficiência com base em parâmetros psicométricos de discriminação, dificuldade e probabilidade acerto casual, permitindo interpretações mais robustas do desempenho do avaliado para além do somatório de acertos [Travitzki 2013,

Andrade et al. 2000]. Todavia, para a aplicação e calibração psicométrica dos itens, é necessária uma quantidade significativa de respondentes, o que torna esse processo operacionalmente complexo e oneroso, especialmente para itens recém-desenvolvidos ou ainda em fase de elaboração [AlKhuyaey et al. 2024].

Nesse sentido, os Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*) vêm obtendo destaque pela incorporação da tecnologia ao campo educacional. Devido à sua escalabilidade, tais modelos apresentam-se como alternativas viáveis para auxiliar em processos educacionais complexos, como a simulação de respondentes em avaliações de múltipla escolha.

Diante dos progressos da Inteligência Artificial aplicada à Educação e à Psicomетria, este trabalho investiga a viabilidade do uso de LLMs na simulação automatizada de respostas. A pesquisa propõe o emprego de engenharia de prompts associada a uma abordagem baseada em agentes para a emulação de respondentes, utilizando questões do ENEM como base experimental. A metodologia consiste em simular o comportamento de estudantes com distintos níveis de proficiência frente a um conjunto selecionado de itens da área de Linguagens e Códigos, visando analisar a coerência psicométrica, a eficiência operacional e o custo computacional das inferências, considerando o impacto de diferentes formatos de serialização de dados.

O objetivo é fornecer subsídios tecnológicos para acelerar o desenvolvimento de bancos de itens e reduzir custos operacionais em diferentes sistemas de avaliação. Para orientar esta investigação, estabeleceram-se as seguintes Questões de Pesquisa:

1. **QP01:** Como a simulação de respostas via LLMs pode reproduzir com fidedignidade o comportamento de respondentes e respeitando os respectivos parâmetros psicométricos observados nas edições do ENEM?
2. **QP02:** Qual o impacto da abordagem na otimização dos custos computacionais e na viabilidade operacional da calibração em larga escala?

2. Trabalhos Relacionados

O emprego de LLMs para tarefas de simulação e estimação aliadas à TRI tem sido objeto de crescente investigação acadêmica [Souza et al. 2025], [Liu et al. 2025, Benedetto et al. 2024]. Pesquisas recentes exploram essa convergência sob diferentes perspectivas, variando desde a automação da calibração psicométrica [Acquaye et al. 2026] até a validação de consistência pedagógica dos instrumentos avaliativos. O mapeamento da literatura permite agrupar essas iniciativas em duas frentes analíticas: a fidelidade da modelagem psicométrica do respondente e as restrições de infraestrutura e custo operacional das simulações.

No âmbito da modelagem psicométrica, o desafio central reside em extrair comportamentos humanos plausíveis de LLMs sob constructos educacionais formais. Lu e Wang et al. [Lu and Wang 2024] investigaram o uso do GPT-4 guiado por perfis de conhecimento discretos, categorizando os estudantes em estados de domínio ou confusão. Buscando maior alinhamento estatístico, [Souza et al. 2025] adotaram uma perspectiva contínua, parametrizando os *prompts* com base no traço latente (θ) da TRI para gerar respondentes artificiais em Ciências Humanas do ENEM. Contudo, vieses estruturais são frequentemente reportados nessa emulação. Liu et al. [Liu et al. 2025] identificaram que as distribuições de proficiência estimada via LLMs tendem a ser mais estreitas que as

reais, ao passo que Benedetto et al. [Benedetto et al. 2024] observaram que modelos com alta acurácia técnica apresentam dificuldades para mimetizar níveis baixos de proficiência. O presente estudo aborda essa vertente ao transpor a emulação para a área de Linguagens e Códigos — cujas características semânticas impõem desafios distintos de interpretação às LLMs — e ao mitigar distorções de cauda por meio do uso de parâmetros populacionais empíricos do ENEM como calibradores do modelo.

Para além dos aspectos estatísticos, a viabilidade de aplicação dessas técnicas em larga escala impõe restrições severas de custo e processamento computacional. Acquaye et al. [Acquaye et al. 2026] simularam uma sala de aula sintética com múltiplos perfis de habilidade usando dez modelos das famílias Gemma, Llama e Qwen, demonstrando correlações expressivas com dados reais, mas sinalizando a alta complexidade de orquestrar múltiplos perfis de erro. O principal gargalo econômico dessa abordagem é evidenciado no trabalho de Song et al. [Song et al. 2026], que propuseram um sistema multiagente integrado à TRI cuja validação experimental limitou-se a apenas 32 estudantes simulados em decorrência dos elevados custos de inferência via APIs de LLMs. Esta pesquisa mitiga tal limitação ao demonstrar que o gargalo financeiro pode ser tratado na camada de estruturação de dados. Ao introduzir o formato de serialização compacto TOON em substituição ao JSON tradicional, viabiliza-se o emprego de amostragens robustas e estatisticamente compatíveis com os pressupostos analíticos da TRI.

Consequentemente, embora a literatura recente consolide a viabilidade teórica da convergência entre LLMs e TRI, as investigações convergem para a necessidade de garantir a integridade psicométrica sem inviabilizar economicamente os fluxos de geração. O presente artigo posiciona-se na resolução concomitante dessas duas lacunas, validando uma estratégia que preserva as propriedades métricas da calibração de itens ao mesmo tempo em que reduz em cerca de 25% o consumo de *tokens*, oferecendo uma alternativa sustentável para sistemas de avaliação em larga escala.

3. Métodos e Desenho Experimental

Esta seção detalha os métodos e infraestrutura empregados na construção do pipeline experimental para a simulação de respondentes em exames de larga escala. O fluxo operacional integra LLMs, engenharia de *prompt* e automação de fluxos, investigando o comportamento comparativo entre o JSON tradicional e a estrutura *Token-Oriented-Object-Notation* (TOON) quanto ao custo computacional e à consistência das respostas. O arranjo experimental fundamenta-se no Modelo Logístico de Três Parâmetros (ML-3P) da TRI e nos níveis cognitivos da Taxonomia de Bloom para a lógica de simulação.

3.1. Coleta e Processamento dos Dados

A base de dados utilizada nesta pesquisa é constituída por itens do ENEM extraídos dos microdados oficiais disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), abrangendo as edições de 2020 a 2023. Para cada item selecionado, mapeou-se o conteúdo textual — composto por enunciado e alternativas de resposta — juntamente com os parâmetros técnicos oficiais de calibração da TRI.

O critério de filtragem delimitou o *corpus* ao domínio de Linguagens, Códigos e suas Tecnologias. Foram excluídas as questões de língua estrangeira e os itens dependentes de elementos visuais, dado o caráter exclusivamente textual da abordagem. A seleção

priorizou a diversidade de níveis de dificuldade com base no parâmetro b do ENEM, que varia no intervalo de $[-3, 3]$. Ao final, selecionaram-se 12 questões representativas, distribuídas de forma equilibrada entre as edições consideradas.

3.2. Simulação de Estudantes

A etapa de simulação de respostas foi conduzida a partir da geração de traços latentes (θ) para um conjunto de respondentes sintéticos, cujas habilidades foram modeladas a partir de uma distribuição gaussiana truncada no intervalo $[-3, 3]$, com média $\mu = 0$ e desvio padrão $\sigma = 1$. Ao todo, foram geradas 1.000 amostras de estudantes, volume populacional que assegura a estabilidade estatística necessária para a calibração sob o modelo analítico ML-3P, conforme preconizado por [Şahin and Anıl 2017].

Para viabilizar a execução experimental e respeitar os limites operacionais de processamento das APIs dos modelos de linguagem, os respondentes foram organizados em 20 estratos independentes, simulando turmas de 50 estudantes cada. A estruturação em lotes mantém a independência entre as simulações e viabiliza o processamento granular dos modelos frente a diferentes perfis de proficiência.

3.3. Construção dos *Prompts*

Neste estudo foram empregados os modelos de linguagem da família DeepSeek (DeepSeek-Flash e DeepSeek-Pro), selecionados em razão de sua eficiência de custo operacional e alto desempenho em tarefas de raciocínio estruturado. A execução foi orquestrada na plataforma n8n em um fluxo *stateless* por lotes: cada chamada de API processa um subgrupo de respondentes de maneira isolada, sem compartilhamento de estado ou memória entre as execuções, o que mitiga riscos de contaminação de contexto e viabiliza a paralelização assíncrona do fluxo. Foram desenvolvidos dois formatos de instruções, denominados **Prompt JSON** e **Prompt TOON**, diferenciando-se exclusivamente pelo formato de serialização dos dados de entrada e de saída. Ambas as abordagens utilizam o contexto do exame e a lista de habilidades (θ) dos estudantes do lote, aplicando estratégias de *role prompting*, *meta-prompting* e *chain-of-thought*.

Em ambos os cenários de *prompt*, o modelo realiza, em etapa prévia à simulação das respostas, uma estimativa do parâmetro de dificuldade da questão por meio de um sistema de pontuação baseado em três dimensões explícitas: o nível cognitivo exigido segundo a Taxonomia de Bloom, a complexidade textual do enunciado e a plausibilidade dos distratores. Esse procedimento mapeia o item a uma faixa de dificuldade no intervalo $[-3, 3]$, compatível com a escala do ML-3P. Com base nessa estimativa, o modelo simula a resposta de cada estudante de forma estocástica, aplicando o parâmetro de acerto casual fixado em $c = 0.20$ — dada a existência de cinco alternativas para todas as questões do exame —, retornando de forma binária o valor 1 para acerto e 0 para erro.

No cenário do **Prompt JSON**, os componentes da questão são serializados no formato JSON padrão, com a saída controlada por um componente de restrição estrutural (*structured output parser*) que impõe um esquema de campos e cardinalidade fixos, assegurando que cada iteração retorne os metadados da questão e os resultados de todos os respondentes do lote de forma conjunta.

No cenário do **Prompt TOON**, a serialização adota o formato compacto proposto no *benchmark* de [TOON 2026]. Projetado para otimizar o consumo de recursos

em estruturas de dados repetitivas, o formato TOON adota uma sintaxe inspirada em YAML na qual listas homogêneas são codificadas em formato tabular compacto, omitindo a repetição de chaves presentes no JSON. Por se tratar de uma sintaxe recente e não canônica, o *system prompt* incorpora instruções explícitas sobre a gramática tabular esperada, detalhando a estrutura de cabeçalhos e linhas. As questões são enviadas ao modelo já serializadas em TOON para que a sintaxe seja processada continuamente ao longo da chamada. Adicionalmente, este formato incorpora a técnica de *few-shot learning* com exemplos para reforçar a conformidade estrutural da saída, contornando a ausência de um *parser* sintático rígido nativo para essa linguagem.

3.4. Classificação de Dificuldade

A classificação de dificuldade dos itens foi estabelecida com base no parâmetro b do ML-3P, adotando-se limiares fixos sobre a escala $[-3, 3]$. Itens com $b < -0.5$ foram definidos como fáceis, itens com $-0.5 \leq b \leq 0.5$ como médios e itens com $b > 0.5$ como difíceis, critérios que orientaram a seleção das 12 questões utilizadas no experimento.

Para a análise dos resultados simulados, os estudantes sintéticos foram igualmente estratificados por faixa de habilidade com base no traço latente, adotando-se os mesmos pontos de corte em ± 0.5 . Adicionalmente, foram definidos limiares estendidos em ± 1.0 e ± 1.5 , permitindo uma segmentação mais granular das faixas de habilidade e possibilitando o monitoramento e a identificação de eventuais vieses ou desvios gerados pelos modelos ao longo de toda a distribuição do traço latente.

4. Resultados

Esta seção apresenta os dados empíricos obtidos nos experimentos de simulação, divididos entre os indicadores de aderência psicométrica e os parâmetros de consumo e custo computacional das inferências estruturadas.

Para avaliar o comportamento dos modelos, as respostas geradas pelos LLMs foram confrontadas com um *benchmark* baseado nos parâmetros oficiais do INEP, calculados via Curva Característica do Item (CCI) do modelo ML-3P. Esse referencial utiliza os três parâmetros reais (a, b, c) de cada questão para computar a probabilidade teórica de acerto $P(\theta)$. As respostas simuladas passaram pelo mesmo processo de estratificação por faixas de habilidade latente (θ). A aderência psicométrica entre as taxas médias de acerto simuladas e as probabilidades teóricas esperadas pela CCI foi quantificada pelos índices de correlação de Pearson (r), erro absoluto médio (MAE) e erro percentual absoluto médio (MAPE), conforme consolidado na Tabela 1.

Os dados da Tabela 1 revelam correlações positivas em todas as configurações testadas, com valores de r variando entre 0,70 e 0,97 no referencial com parâmetros oficiais. Nos itens fáceis e médios, o DeepSeek-Flash e o DeepSeek-Pro apresentaram alta aderência às trajetórias previstas pelas CCIs.

Contudo, observou-se uma tendência sistemática de superestimação das taxas de acerto em ambos os modelos nos itens de alta dificuldade ($b > 0,5$), na qual respondentes sintéticos de baixa proficiência obtiveram taxas de acerto significativamente superiores às projetadas matematicamente pela TRI. Esse viés foi mais acentuado nos modelos Flash, enquanto o Pro com formato TOON demonstrou maior consistência nesse estrato em todos os limiares avaliados.

Tabela 1. Métricas de aderência psicométrica por nível de dificuldade do item, intervalo de θ e configuração modelo e formato de serialização.

Intervalo	Item	Flash-V4 / JSON			Flash-V4 / TOON			Pro-V4 / JSON			Pro-V4 / TOON		
		<i>r</i>	MAE	MAPE	<i>r</i>	MAE	MAPE	<i>r</i>	MAE	MAPE	<i>r</i>	MAE	MAPE
$\theta = \pm 0,5$	Fácil	0,924	0,142	22,51%	0,771	0,138	25,45%	0,782	0,113	25,18%	0,810	0,118	25,15%
	Médio	0,939	0,075	17,72%	0,797	0,156	44,26%	0,850	0,143	38,06%	0,854	0,128	30,92%
	Difícil	0,859	0,271	107,66%	0,738	0,273	109,63%	0,706	0,252	97,39%	0,841	0,227	82,10%
$\theta = \pm 1,0$	Fácil	0,955	0,117	19,42%	0,868	0,132	32,66%	0,835	0,112	36,34%	0,886	0,109	28,56%
	Médio	0,973	0,057	15,08%	0,894	0,136	41,66%	0,925	0,115	33,98%	0,917	0,095	22,44%
	Difícil	0,844	0,237	87,19%	0,758	0,239	90,55%	0,740	0,222	82,67%	0,869	0,197	62,12%
$\theta = \pm 1,5$	Fácil	0,963	0,110	29,44%	0,889	0,146	57,73%	0,902	0,104	51,61%	0,917	0,110	42,87%
	Médio	0,979	0,053	14,43%	0,925	0,120	39,29%	0,953	0,090	27,18%	0,943	0,087	21,85%
	Difícil	0,863	0,188	68,83%	0,791	0,167	58,31%	0,828	0,176	60,90%	0,832	0,162	52,95%

Ademais, o modelo DeepSeek-Flash apresentou inversão na ordenação de dificuldade em estratos idênticos de proficiência, gerando taxas de acerto simuladas mais elevadas para itens difíceis do que para itens fáceis — superestimação que constitui a causa direta dos valores de MAPE elevados observados nesse estrato na Tabela 1. O modelo DeepSeek-Pro, por sua vez, demonstrou maior regularidade na preservação da ordenação monotônica preconizada pela TRI. Desvios residuais persistiram em todos os limiares avaliados, evidenciando que o viés de superestimação nos itens difíceis foi suavizado, embora ainda permaneça observável.

Na dimensão da eficiência computacional, a Tabela 2 consolida o consumo de *tokens* de entrada (*prompt*) e saída (*completion*), o coeficiente de variação (CV) e os custos financeiros indexados às APIs para os quatro cenários experimentais.

Tabela 2. Consumo de tokens e custo por modelo e formato de serialização.

Modelo	Formato	Tokens de Entrada			Tokens de Saída			Total Geral	Δ %	Custo (US\$)
		Total	Média	CV (%)	Total	Média	CV (%)			
Flash V4	JSON	582.128	2.426	4,98	1.670.754	6.961	29,15	2.252.882	base	0,7116
	TOON	435.328	1.814	6,75	1.791.943	7.466	28,39	2.227.271	-1,14	0,5618
Pro V4	JSON	577.848	2.408	5,01	2.592.206	10.801	33,99	3.170.054	base	10,0263
	TOON	438.448	1.827	6,64	1.950.629	8.128	35,96	2.389.077	-24,64	7,5511
Média geral		508.438	2.119	—	2.001.383	8.339	—	2.509.821	—	—

A transição do formato JSON para o formato TOON resultou em uma redução estável de aproximadamente 25% no volume de *tokens* de entrada em ambos os modelos. No entanto, o volume de *tokens* de saída variou conforme a arquitetura do modelo. No DeepSeek-Flash, o ganho na entrada foi atenuado por um incremento no volume de geração do *completion*. Já o DeepSeek-Pro apresentou aderência estrita à gramática tabular do formato TOON, reduzindo o consumo em ambas as dimensões de tráfego. Essa otimização gerou uma economia líquida de aproximadamente 781.000 *tokens* em comparação com o cenário JSON, uma redução de 24,64% no volume total de dados.

5. Discussão

A análise concomitante dos dados psicométricos e computacionais permite avaliar o alcance e as limitações da metodologia proposta, respondendo diretamente às questões de pesquisa estabelecidas.

5.1. Análise da Fidelidade Psicométrica (QP01)

A primeira questão de pesquisa (QP01) investigava em que medida a simulação via LLMs reproduz com fidelidade o comportamento de respondentes humanos e os parâmetros do ENEM. Os coeficientes de correlação de Pearson obtidos validam a inclusão de critérios baseados na Taxonomia de Bloom e na complexidade textual como ancoragens cognitivas nos *prompts*. Contudo, o comportamento das métricas revela que essa fidelidade é altamente sensível à granularidade da estratificação dos respondentes e ao nível de dificuldade do item.

Os dados demonstram que, ao adotar um intervalo restrito de θ ($\pm 0,5$), os modelos apresentam menor aderência e maior erro absoluto, efeito pronunciado nos itens de alta complexidade. Embora a magnitude do erro absoluto (MAE) e do percentual (MAPE) seja elevada nesses estratos, esse comportamento é justificado por um efeito de escala: como as probabilidades reais de acerto de respondentes de baixa proficiência em itens difíceis são assintoticamente próximas a zero nos parâmetros do ENEM, qualquer ruído estocástico ou acerto casual do LLM infla drasticamente as métricas de erro relativo. Contudo, os altos índices de correlação de Pearson obtidos provam que, apesar dessa distorção de escala nos extremos inferiores, o modelo replica com precisão a tendência de acerto e a trajetória de crescimento das curvas teóricas.

Por outro lado, quando se flexibiliza o intervalo do traço latente para faixas estendidas em $\pm 1,0$ e $\pm 1,5$, observa-se uma melhora progressiva nas métricas. Essa expansão de escopo permite que os modelos processem uma transição de habilidade mais acentuada, atenuando o peso da superestimação local e fazendo com que as taxas simuladas converjam para as trajetórias teóricas das CCIs oficiais, conforme documentado por Benedetto et al. [Benedetto et al. 2024].

Essa dinâmica de convergência é condicionada pela arquitetura subjacente. No modelo probabilístico convencional DeepSeek-Flash, a flexibilização melhora os índices globais, mas não elimina o viés de inversão de ordenação nos estratos mais baixos de proficiência, onde itens difíceis registraram mais acertos que itens fáceis. Esse viés alinha-se ao comportamento identificado por Acquaye et al. [Acquaye et al. 2026], que apontam a tendência de modelos de menor capacidade em projetar o *output* a partir de sua própria proficiência resolutiva, falhando na emulação de perfis de erro. A distorção foi mitigada pelo DeepSeek-Pro (especialmente no formato TOON), que preservou a monotonicidade teórica da TRI. Esse comportamento corrobora as conclusões de Liu et al. [Liu et al. 2025] sobre a necessidade de modelos estruturalmente robustos para emular com fidelidade distribuições psicométricas complexas.

5.2. Análise do Impacto Operacional e de Custo (QP02)

A segunda questão de pesquisa (QP02) avaliava o impacto da abordagem na otimização dos custos computacionais e na viabilidade operacional da calibração em larga escala. Os resultados empíricos consolidam que a escolha do formato de serialização de dados atua como variável crítica para a sustentabilidade econômica do processo, contanto que o comportamento sintático do modelo seja considerado. A redução de aproximadamente 25% nos *tokens* de entrada pelo formato TOON, em ambos os modelos, valida a hipótese de eficiência decorrente da eliminação de chaves e metadados redundantes nativos do padrão JSON, em conformidade com a especificação estrutural proposta por [TOON 2026].

Contudo, a análise detalhada dos *tokens* de saída (*completion*) revela que a viabilidade econômica do formato está atada à capacidade arquitetural do modelo. No modelo probabilístico convencional DeepSeek-Flash, a economia gerada no *prompt* foi neutralizada por um incremento volumétrico na saída. Os dados brutos indicam que esse modelo apresentou falhas na aderência à gramática tabular não canônica do TOON, gerando caracteres de preenchimento ou repetições para alcançar a conformidade, o que elevou o custo de processamento final. Por outro lado, o modelo de raciocínio analítico DeepSeek-Pro demonstrou alta consistência sintática, reduzindo o tráfego em ambas as dimensões e consolidando a economia total de 24,64% (781.000 *tokens* poupados).

O cruzamento desses dados de infraestrutura com os indicadores psicométricos da Tabela 1 confirma que a compressão de dados do formato TOON não causou degradação estatística nos resultados. O erro absoluto médio (MAE) manteve-se estável e os índices de correlação preservaram patamares equivalentes aos do JSON sob a mesma arquitetura de raciocínio, com o DeepSeek-Pro registrando os menores valores de MAE nos itens difíceis em todos os limiares avaliados. Esse equilíbrio supera o gargalo econômico observado em simulações educacionais baseadas em agentes, como a de Song et al. [Song et al. 2026], cujo alcance limitou-se a apenas 32 respondentes. A combinação do formato TOON com modelos estruturados para cadeias de pensamento prova ser operacionalmente viável para processar amostragens estatisticamente robustas ($N = 1.000$), mitigando o custo por inferência sem comprometer a integridade da calibração psicométrica.

6. Considerações Finais

Este trabalho avaliou o uso de LLMs para simular o comportamento de respondentes no ENEM. O grande diferencial metodológico da abordagem foi integrar engenharia de *prompt*, o modelo ML-3P da TRI e a Taxonomia de Bloom para fazer com que os modelos estimassem a dificuldade das questões e gerassem respostas estocásticas sem depender ou ter acesso prévio aos parâmetros oficiais de calibração do exame.

Os resultados mostram que a fidelidade dessa simulação varia conforme a arquitetura do modelo. Modelos probabilísticos convencionais apresentaram limitações na emulação de erros e inverteram a ordenação de dificuldade dos itens. Em contrapartida, modelos baseados em cadeias de raciocínio mitigaram esses vieses e preservaram a monotonicidade teórica da TRI. Além disso, o formato TOON confirmou que é possível otimizar a estrutura de dados e reduzir o consumo de *tokens* sem comprometer a aderência psicométrica, evidenciando que ganhos de qualidade e eficiência podem ser obtidos de forma complementar.

Cabe destacar as limitações do presente estudo. A validação restringiu-se a 12 itens do ENEM, concentrados na área de Linguagens e Códigos e avaliados exclusivamente por modelos da família DeepSeek. A quantidade permitiu contemplar os três níveis de dificuldade, considerando também as restrições operacionais do desenho experimental.

Para trabalhos futuros, propõe-se ampliar o *corpus* do ENEM e aplicar o fluxo em Ciências da Natureza e Matemática. É relevante testar estratégias de *prompting* para modelar os padrões de erro humanos e avaliar o desempenho do formato TOON em modelos de código aberto, e investigar formatos emergentes de serialização como alternativas para a otimização do consumo em LLMs. A metodologia mostrou-se viável para apoiar a pré-calibração psicométrica e auxiliar na criação de bancos de itens de forma mais acessível.

Cuidados Éticos

Esta pesquisa utiliza dados secundários de acesso público e anonimizados, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Por operar estritamente sobre bases de dados abertas, descaracterizadas na origem e amparadas pela Lei de Acesso à Informação (LAI), o estudo prescinde de submissão a Comitê de Ética em Pesquisa, cumprindo integralmente as diretrizes da LGPD e os devidos cuidados éticos.

Agradecimentos

O presente trabalho foi financiado parcialmente pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Os autores agradecem ao Observatório de Dados da Graduação (ODG) e à UFRPE pelo suporte institucional.

Declaração sobre o Uso de Inteligência Artificial

Para além do escopo metodológico explicitado nos experimentos, ferramentas de Inteligência Artificial Generativa foram utilizadas estritamente no auxílio à redação, refinamento gramatical e organização textual deste documento. Os autores revisaram integralmente o texto gerado e assumem total responsabilidade científica e editorial pelo conteúdo final.

Referências

- Acquaye, C. et al. (2026). Take out your calculators: Estimating the real difficulty of question items with llm student simulations. *arXiv preprint arXiv:2601.09953*.
- AlKhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Andrade, D. F. d., Tavares, H. R., and Valle, R. d. C. (2000). *Teoria da Resposta ao Item: Conceitos e Aplicações*. Associação Brasileira de Estatística, São Paulo.
- Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., and Buttery, P. (2024). Using llms to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.
- Liu, Y., Bhandari, S., and Pardos, Z. A. (2025). Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.
- 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 '24*, pages 16–27, New York, NY, USA. Association for Computing Machinery.
- Pasquali, L. (2003). Os itens de um teste. In *Psicometria: teoria dos testes na psicologia e na educação*, pages 137–160. Vozes, Petrópolis.
- Sobrinho, A. and Cavalcanti, C. R. (2023). Políticas educacionais de avaliação em larga escala no brasil: o enem em foco. *Revista Contemporânea de Educação*, 18(41):226–246.

- Song, Y., Li, H., and Zheng, Q. (2026). Multi-agent automated item generation based on large language models. *Journal of Educational Technology & Society*, 29(2):233–248.
- Souza, J. A. d., Rodrigues, E., Gouveia, R., and Alves, G. (2025). Uma abordagem baseada em simulação com grandes modelos de linguagem para validação de dificuldade de questões do enem a partir da tri. In *Anais do Workshop de Aplicações Práticas de Learning Analytics e Inteligência Artificial no Brasil (WAPLA)*, pages 91–99, Porto Alegre. Sociedade Brasileira de Computação.
- TOON (2026). Benchmarks. Online. Disponível em: <https://toonformat.dev/guide/benchmarks.html>. Acesso em: 24 abr. 2026.
- Travitzki, R. (2013). Enem: limites e possibilidades do exame nacional do ensino médio após a adoção da tri. *Estudos em Avaliação Educacional*, 28(67):44–76.
- Şahin, A. and Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, 17(1):321–335.