Uma Abordagem Baseada em Simulação com Grandes Modelos de Linguagem para Validação de Dificuldade de Questões do ENEM a partir da TRI

Jéssica Alves de Souza, Ebony Rodrigues, Roberta Gouveia, Gabriel Alves

[1]Departamento de Estatística e Informática (DEINFO)
Universidade Federal Rural de Pernambuco (UFRPE)

{jessica.alvess,ebony.marquesr,roberta.gouveia,gabriel.alves}@ufrpe.br

Abstract. *The creation of assessment items is a complex task, especially in defining their difficulty level. This study proposes the use of Large Language Models (LLMs) to estimate the difficulty parameter b of the Item Response Theory (IRT) for ENEM questions. The LLMs simulate student responses at different ability levels, allowing the calculation of b and comparison with the known parameters of ENEM items. Different prompts were tested, and one of them showed promising results by estimating b solely based on the question text and answer options. Error metrics indicated minor discrepancies, without compromising the classification into difficulty levels, achieving over 50% accuracy across the three levels. The analysis demonstrated that this strategy is feasible for automated difficulty assessment.*

Resumo. *A criação de questões avaliativas é uma tarefa complexa, especialmente na definição do nível de dificuldade. Este trabalho propõe o uso de Grandes Modelos de Linguagem (LLMs) para estimar o parâmetro de dificuldade b da Teoria de Resposta ao Item (TRI) em questões do ENEM. Os LLMs simulam respostas de estudantes com diferentes habilidades, permitindo calcular b e comparar com os parâmetros conhecidos de questões do ENEM. Diferentes prompts foram testados, e um deles apresentou resultados promissores estimando b apenas com base no texto e alternativas da questão. As métricas de erro indicaram pequenas discrepâncias, sem comprometer a classificação em níveis de dificuldade, com acurácia acima dos 50%, para os 3 níveis de dificuldade. A análise mostrou que a estratégia é viável para avaliação automatizada da dificuldade.*

1. Introdução

A criação de questões avaliativas é uma tarefa complexa, que exige tempo, conhecimento pedagógico e cuidado na definição, entre outras variáveis, do nível de dificuldade de cada item. A dificuldade em estimar com precisão esse nível pode gerar problemas significativos, como a elaboração de itens excessivamente fáceis ou difíceis, vieses não intencionais e falta de consistência entre diferentes provas, impactando a validade das avaliações [Kurdi et al. 2020, Mulla and Gharpure 2023]. Nesse contexto, a utilização de Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*) pode contribuir para a eficiência e a escalabilidade do processo de análise de questões a fim de determinar a sua dificuldade.

DOI: 10.5753/wapla.2025.16245

A Teoria de Resposta ao Item (TRI) fornece um arcabouço estatístico robusto para estimar parâmetros que descrevem o desempenho esperado dos respondentes [Baker 2001]. Em particular, o parâmetro b, que indica o nível de habilidade no qual a probabilidade de acerto é de 50%, constitui um indicador quantitativo da dificuldade do item. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza publicamente os parâmetros da TRI das questões aplicadas no Exame Nacional do Ensino Médio (ENEM), oferecendo uma referência para calibrar métodos automatizados de estimativa de dificuldade nessas questões [Tomikawa and Uto 2024].

Este trabalho propõe a utilização de LLMs para estimar a dificuldade das questões, empregando as questões do ENEM como base para a validação. Para esse fim, os LLMs irão simular respostas de estudantes com diferentes níveis de habilidades e diferentes *prompts*, a fim de responder a algumas questões da área de ciências humanas do ENEM de 2023. O parâmetro de dificuldade b da TRI pode então ser calculado com base nas respostas para cada questão, aferindo-se as métricas de erro em relação ao parâmetro b original da questão. Esse processo auxilia a elaboração dos *prompts* mais eficazes na definição da dificuldade das questões. Assim, este trabalho se propõe a responder às seguintes perguntas de pesquisa:

1. Como os Grandes Modelos de Linguagem podem ser utilizados para estimar o parâmetro b da TRI em questões do ENEM a partir da simulação de respostas de estudantes?
2. Em que medida essas simulações calibradas com os parâmetros TRI fornecidos pelo INEP conseguem reproduzir a dificuldade observada nas questões reais?

2. Trabalhos Relacionados

Alguns trabalhos já buscaram avaliar o uso de grandes modelos de linguagem (LLMs) ou outras técnicas para estimar a dificuldade em diferentes contextos e níveis educacionais [Liu et al. 2025, Tomikawa and Uto 2024, Marinho et al. 2023]. Apesar de alguns trabalhos focarem em diferentes áreas do conhecimento [Liu et al. 2025], muitos focam habilidades ligadas a linguagens, como a habilidade de leitura [Tomikawa and Uto 2024, Jain et al. 2025].

A estimativa dos parâmetros da TRI também já foi avaliada em alguns trabalhos [Liu et al. 2025, Jain et al. 2025]. A utilização de respostas geradas sinteticamente com diferentes LLMs e as respostas de estudantes foram utilizadas para estimar os parâmetros do TRI, obtendo uma alta correlação (0,87 a 0,93). Apesar da alta correlação, os LLMs apresentam uma distribuição de proficiência mais estreita, o que limita sua capacidade de representar a variabilidade humana [Liu et al. 2025].

O sistema KAQG, proposto por [Chen and Shiu 2025], cria automaticamente questões de prova com dificuldade controlada usando grafos de conhecimento e geração aumentada por recuperação (RAG). Primeiro, o material didático é transformado em um grafo no qual entidades e relações são bem estruturadas. Depois, o componente de RAG faz buscas nesse grafo para encontrar fatos relevantes e os envia para o modelo de linguagem. Com esses dados, o modelo gera perguntas ajustadas pelos critérios da Teoria da Resposta ao Item (dificuldade, discriminação e chance de chute) e pelos níveis de Bloom. Nos testes, o TRI de cada questão produzida correspondeu aos *benchmarks* de especialistas. Entre os ganhos estão o controle preciso dos parâmetros, a integração automática

de conhecimento externo via RAG e a capacidade de gerar muitas perguntas de forma escalável.

O modelo logístico de três parâmetros (ML-3P) para estimar erros padrão (EP) de itens de um teste real de matemática foi utilizado para investigar a máxima verossimilhança marginal [Ogbonna and Opara 2018]. Assim, diferentes medidas de erro foram utilizadas para a seleção de itens com melhor estabilidade.

O uso de *prompt engineering* também ocorreu para simular respostas de estudantes com diferentes níveis de proficiência, utilizando o GPT-3.5 [Benedetto et al. 2024] . A abordagem envolveu a criação de um *reference prompt* (RP) para gerar saídas padronizadas, que foram testadas em múltiplos conjuntos de dados e avaliadas por monotonicidade da acurácia e aderência à dificuldade das questões. Embora o método tenha se mostrado promissor, os autores observaram instabilidade nas estimativas de dificuldade e limitações na generalização para outros modelos.

De modo geral, os trabalhos analisados demonstram o potencial dos LLMs no suporte à educação, seja por meio da geração de conteúdo, simulação de respostas ou avaliação automatizada. Contudo, muitos deles se concentram em análises qualitativas ou validações parciais. O presente estudo se distingue ao propor uma abordagem *end-to-end* que integra a simulação de respostas com perfis variados de habilidade e validação estatística por meio da Teoria da Resposta ao Item, utilizando métricas quantitativas para garantir a consistência e utilidade dos itens produzidos.

3. Metodologia

Esta seção descreve a metodologia utilizada no desenvolvimento deste trabalho. A solução emprega LLMs e a Teoria da Resposta ao Item (TRI) a fim de aferir o nível de dificuldade de questões do ENEM. Esse tipo de solução é importante para tornar escalável a definição da dificuldade de itens avaliativos em soluções educacionais de larga escala, como em Cursos Online Abertos e Massivos (MOOCs, do inglês *Massive Open Online Courses*), em que a definição da dificuldade exclusivamente por pedagogos e educadores se torna inviável.

O processo de identificação da dificuldade das questões inicia-se com a definição do *prompt* que será enviado à LLM junto com a questão e suas opções de resposta. Então, são geradas respostas simuladas para estudantes com diferentes graus de habilidade, estimando-se o parâmetro b da TRI, a fim de descobrir o nível de dificuldade da questão. Neste trabalho foi utilizado o *Modelo Logístico de 3 Parâmetros* (ML-3P, ou 3PL *model*, do inglês *Three-Parameter Logistic Model*) para o TRI, uma vez que é o modelo utilizado pelo INEP, na disponibilização dos dados do ENEM.

3.1. Construção dos *Prompts*

Para este estudo, foram selecionadas 20 questões de Ciências Humanas do ENEM 2023 (caderno azul, códigos 47 a 68), previamente filtradas para excluir itens com imagens. Essas questões disponibilizadas pelo INEP *online* possuem os valores dos parâmetros a, b e c do TRI. O parâmetro b foi usado como valor de referência de cada questão selecionada. Assim, para cada questão, foi realizada a simulação da resposta de 1000 alunos, quantidade recomendada pelo TRI [Andrade et al. 2000], com habilidades entre -3 e 3,

seguindo uma distribuição gaussiana. Portanto, a maior parte dos alunos possui habilidade 0, indicando que possui uma chance de 50% de acertar uma questão cujo parâmetro de dificuldade b também seja 0.

O LLM utilizado foi o DeepSeek-V3-0324, dado o seu baixo custo e bom desempenho. Para a simulação de respostas para 1000 alunos em 20 questões, observaram-se falhas de execução atribuídas ao limite de *tokens* por chamada da API. Portanto, foi realizado o particionamento em 50 chamadas assíncronas, cada uma processando 20 estudantes de habilidades diferentes, o que garantiu respostas completas sem erros de quantidade de *tokens*, mantendo uma performance escalável para a solução.

Foram utilizados três *prompts* distintos para a simulação de respostas, para os quais foi informado o contexto de resolução de questões do ENEM e a habilidade (θ) dos estudantes. Todos os *prompts* consideram o Modelo Logístico de 3 Parâmetros (ML-3P) da TRI, em que a, b e c são os parâmetros de discriminação, de dificuldade e de acerto casual, respectivamente. Esse trabalho foca especificamente o parâmetro de dificuldade b do referido modelo, com valores variando de -3 a 3.

No $Prompt\ A$, a habilidade θ de cada estudante é definida de probabilisticamente, seguindo uma distribuição gaussiana, em conformidade com a TRI. O parâmetro de dificuldade b é definido de forma aleatória pela própria LLM e adota-se um valor fixo de 20% para o parâmetro c, de acerto casual, dado que existem 5 opções por questão. Assim, as respostas dos estudantes são geradas de forma probabilística, retornando 1 caso tenha sido correta e 0 caso tenha sido incorreta.

O $Prompt\ B$ oferece um contexto mais específico, voltado para questões de Ciências Humanas do ENEM, informando ainda os parâmetros a, b e c originais da questão. Já a habilidade θ dos estudantes é definida seguindo uma distribuição gaussiana, conforme a TRI. Vale salientar que esse *prompt* não seria aplicável um muitos casos, por necessitar dos parâmetros do TRI, que precisam ter sido obtidos anteriormente, seja pela resposta direta dos alunos, seja pela estimativa dos parâmetros. Importa destacar que este *prompt* não tem como objetivo estimar parâmetros desconhecidos, mas sim funcionar como um *baseline*, juntamente com o $Prompt\ A$, servindo de base de comparação para os resultados das simulações obtidas em outros *prompts*, como o $Prompt\ C$.

O $Prompt\ C$, por fim, retorna a resposta dos estudantes com base em regras explicitadas no próprio *prompt*, no texto e opções da questão e na respostas dos estudantes com base em uma habilidade θ gerada seguindo uma distribuição gaussiana. O *prompt* inclui erros sistemáticos como 40% de chance de escolher um distrator plausível e 15% de chance de escolher uma opção absurda, além de um viés cognitivo de preferência por alternativas centrais B, C, ou D. A LLM é estão instruída a estimar a dificuldade da questão (parâmetro b do ML-3P), considerando uma escala contínua de -3.0 (nível mínimo de dificuldade) a 3.0 (nível mínimo de dificuldade). A LLM é então instada a escolher uma opção da questão a partir da habilidade do estudante, que varia entre -3.0 (defasagem grave) a 3.0 (excelência) nos conhecimentos esperados para o ENEM e a dificuldade estimada da questão. O *prompt* então retorna um JSON, com uma lista de escolhas (A, B, C, D e E), uma lista binária de acertos (0/1), o percentual de acerto de cada questão e a dificuldade calculada. A lista binária de acertos e os parâmetros de habilidade θ são utilizados para calcular os parâmetros do ML-3P, enquanto os demais atributos da saída

podem ser utilizados no ajuste posterior do *prompt*.

3.2. Validação Estatística dos Parâmetros TRI

A fim de responder à primeira pergunta de pesquisa, para avaliar a capacidade do simulador de estimar os parâmetros a, b e c, foram geradas estimativas com base no $ML - 3P$. Os parâmetros a, b e c foram estimados a partir da média dos limites inferior e superior do intervalo de confiança para cada questão, considerando os *prompts* A, B e C. O parâmetro b_0 de cada questão equivalente dos microdados do ENEM 2023 foi utilizado como valor de referência.

Com base nos erros obtidos entre os valores estimados e os reais, foram calculadas quatro métricas estatísticas sobre o conjunto das questões selecionadas: erro médio de viés (MBE), erro absoluto médio (MAE), erro quadrático médio (MSE) e erro percentual médio (MAPE) [Hyndman and Athanasopoulos 2018].

Uma vez que este trabalho trata da simulação de respostas com base em probabilidades, é interessante que a validação seja realizada com base no intervalo de confiança (IC). Neste trabalho utilizou-se o intervalo de confiança de 95%, em que os limites inferior e superior definem a faixa na qual se espera que o valor verdadeiro do parâmetro esteja contido em aproximadamente 95% das amostragens, assumindo que o mesmo procedimento de coleta e estimativa seja repetido sob as mesmas condições [Bussab and Morettin 2017].

Assim, se o valor original do parâmetro estiver dentro do intervalo de confiança calculado a partir da simulação, considera-se que o modelo estimou corretamente o parâmetro. Contudo, mesmo que o valor original do b não esteja dentro do IC 95%, um valor próximo já indica que o modelo de classificação de dificuldade pode ser ajustado a partir da adequação dos limites adotados para a dificuldade média da questão.

Para estimar a variabilidade dos parâmetros gerados pelo modelo TRI, foi empregada a técnica de reamostragem *bootstrap*, com 1000 amostras geradas com reposição a partir do conjunto original de respostas dos estudantes [Hambleton et al. 1991]. Para cada amostra, os parâmetros a, b e c foram reestimados para todas as questões, totalizando 1000 repetições do processo. Vale ressaltar que, apesar do cálculo de todos os parâmetros do TRI, dado o uso do ML-3P, apenas o parâmetro b foi utilizado neste trabalho.

Com base nas distribuições obtidas para cada parâmetro, foram calculados os intervalos de confiança de 95% utilizando a distribuição t de Student, por meio do ambiente Google Colab, com a linguagem Python e a biblioteca `scipy`.

3.3. Classificação dos Níveis de Dificuldade

A segunda pergunta de pesquisa é respondida a partir da classificação das perguntas nos níveis fácil, médio ou difícil. Esse nível de dificuldade é definido a partir do parâmetro b do modelo do $ML - 3P$. Assim, para $-0,5 \leq b \leq 0,5$, considera-se a questão de nível médio, e para $b < -0,5$ e $b > 0,5$, a questão é considerada fácil e difícil, respectivamente. Esses intervalos foram utilizados para calcular o nível de dificuldade original de cada questão com base no parâmetro b da questão nos dados do INEP.

Para estimar a dificuldade a partir dos dados simulados no modelo, foi realizada uma flexibilização desses limites, a fim de prover uma correção de vieses do modelo.

Nesse caso, foram testados também os limites $-0,75 \leq b \leq 0,75$ e $-1,0 \leq b \leq 1,0$ para que uma questão seja considerada de dificuldade média. Após a conversão do parâmetro b no nível de dificuldade, é realizado o *ordinal encoding* da variável, sendo 1 a dificuldade fácil, 2, a dificuldade média e 3, a difícil. Assim, foram calculadas as métricas de erro e a acurácia de cada um dos *prompts*, adotando diferentes limites para a discretização.

4. Limitações

Este trabalho foi conduzido exclusivamente com o modelo DeepSeek-V3-0324, sem realizar comparações com outros LLMs, o que restringe a análise da generalização. Não foram empregadas técnicas de *fine-tuning*, *soft prompting*, nem mecanismos de Retrieval-Augmented Generation (RAG) para incorporar conhecimentos específicos do ENEM, potencialmente limitando a acurácia das estimativas do parâmetro b. Ademais, a utilização de um conjunto reduzido de questões de uma área de conhecimento específica para o ENEM de 2023 também pode impactar a robustez estatística, dificultando a extrapolação dos resultados para um universo mais amplo de itens.

5. Resultados e Discussões

A seguir, são apresentados os resultados obtidos e as discussões desses resultados com o objetivo de responder às perguntas de pesquisa definidas para este trabalho. Assim, a Seção 5.1 apresenta os resultados para a primeira pergunta de pesquisa, enquanto a Seção 5.2 foca os resultados da segunda pergunta de pesquisa.

5.1. Como os Grandes Modelos de Linguagem podem ser utilizados para estimar o parâmetro *b* da TRI em questões do ENEM a partir da simulação de respostas de estudantes?

A Tabela 1 apresenta a média e o intervalo de confiança aferidos para os diferentes *prompts*, para cada uma das 20 questões selecionadas. Apesar de os valores de referência estarem fora do IC 95% em praticamente todos os casos, vale ressaltar que em muitos casos o valor de referência está próximo do valor encontrado.

Já a Tabela 2 apresenta as métricas de erro calculadas para cada um dos *prompts*. Um primeiro ponto a ser observado é que não seria recomendado tomar decisões com base na métrica $MAPE$, pois esse contexto lida com valores muito pequenos, resultando em grandes variações nessa métrica.

Já a métrica MBE aponta que o $Prompt\ B$ teve uma tendência a superestimar um pouco a dificuldade das questões, enquanto os *prompts* A e C tiveram uma tendência a subestimar a dificuldade das questões. Já as métricas MAE e MSE mostram que o $Prompt\ B$ obteve os melhores resultados. Por sua vez, o $Prompt\ C$ teve uma tendência a subestimar bastante a dificuldade de alguns itens.

Apesar de as métricas do erro apontarem para algumas discrepâncias, em muitos casos essa discrepância não foi tão grande, o que viabilizaria a abordagem proposta. Contudo, seria importante experimentar novos *prompts*, novas questões e áreas de conhecimento, a fim de observar se esse comportamento seria similar. Outras técnicas como o *prompt tuning* ou o uso de outras LLMs também poderia contribuir para uma melhoria nos resultados.

Tabela 1. Média e intervalo de confiança esperados do parâmetro b para os *prompts* A, B e C.

♯	b_0	$b_A\,[IC95\%]$	$b_B\,[IC95\%]$	$b_C\,[IC95\%]$
1	0,006	0,238 [0,237; 0,240]	0,577 [0,575; 0,579]	-0,175 [-0,181; -0,168]
2	0,949	0,403 [0,398; 0,408]	0,761 [0,759; 0,763]	-0,020 [-0,032; -0,008]
3	0,671	0,315 [0,318; 0,317]	0,529 [0,527; 0,531]	-0,391 [-0,403; -0,379]
4	1,191	0,383 [0,373; 0,393]	0,793 [0,790; 0,795]	-0,324 [-0,335; -0,314]
5	1,299	0,346 [0,343; 0,349]	1,260 [1,256; 1,263]	3,000 [3,000; 3,000]
6	1,185	0,281 [0,277; 0,286]	0,648 [0,646; 0,651]	-0,510 [-0,524; -0,496]
7	0,821	0,312 [0,310; 0,314]	0,541 [0,539; 0,543]	-0,587 [-0,602; -0,571]
8	2,638	0,760 [0,753; 0,767]	1,405 [1,401; 1,410]	1,689 [1,676; 1,702]
9	0,673	0,264 [0,262; 0,266]	0,442 [0,440; 0,444]	-0,690 [-0,715; -0,665]
10	0,066	0,497 [0,493; 0,501]	0,702 [0,700; 0,704]	-0,419 [-0,432; -0,406]
11	0,245	0,277 [0,274; 0,280]	0,580 [0,578; 0,582]	-0,506 [-0,523; -0,489]
12	0,115	0,553 [0,547; 0,559]	0,704 [0,701; 0,706]	-0,413 [-0,430; -0,395]
13	0,086	0,290 [0,288; 0,292]	0,592 [0,590; 0,595]	-0,829 [-0,850; -0,808]
14	-0,020	0,467 [0,460; 0,474]	0,698 [0,696; 0,700]	0,101 [0,090; 0,111]
15	0,393	0,321 [0,319; 0,323]	0,661 [0,659; 0,664]	2,945 [2,933; 2,956]
16	0,709	0,568 [0,562; 0,574]	0,679 [0,677; 0,681]	-0,477 [-0,483; -0,471]
17	0,318	0,315 [0,312; 0,318]	0,581 [0,579; 0,582]	-0,494 [-0,509; -0,479]
18	-0,620	0,467 [0,462; 0,472]	0,697 [0,695; 0,699]	-0,963 [-0,986; -0,940]
19	0,129	0,262 [0,259; 0,264]	0,579 [0,577; 0,581]	-0,819 [-0,843; -0,796]
20	0,081	0,730 [0,723; 0,736]	0,716 [0,713; 0,718]	0,021 [-0,005; 0,046]

Tabela 2. Métricas de erro do parâmetro b para os *prompts* A, B e C.

♯	*Prompt A*				*Prompt B*				*Prompt C*			
	MAE	MBE	MSE	MAPE	MAE	MBE	MSE	MAPE	MAE	MBE	MSE	MAPE
1	0,232	-0,232	0,054	3831,766	0,571	-0,571	0,326	9425,000	0,181	0,181	0,033	2981,436
2	0,546	0,546	0,298	57,518	0,188	0,188	0,035	19,810	0,969	0,969	0,939	102,130
3	0,356	0,356	0,127	53,045	0,142	0,142	0,020	21,174	1,062	1,062	1,128	158,286
4	0,808	0,808	0,653	67,829	0,399	0,399	0,159	33,452	1,516	1,516	2,298	127,240
5	0,953	0,953	0,908	73,370	0,039	0,039	0,002	2,992	1,701	-1,701	2,895	131,014
6	0,904	0,904	0,817	76,252	0,537	0,537	0,288	45,289	1,695	1,695	2,874	143,010
7	0,509	0,509	0,259	61,980	0,280	0,280	0,078	34,082	1,407	1,407	1,981	171,456
8	1,878	1,878	3,525	71,181	1,232	1,232	1,519	46,720	0,949	0,949	0,900	35,959
9	0,409	0,409	0,167	60,763	0,231	0,231	0,053	34,265	1,363	1,363	1,858	202,574
10	0,430	-0,430	0,185	647,539	0,635	-0,635	0,404	956,511	0,485	0,485	0,236	730,528
11	0,032	-0,032	0,001	12,928	0,334	-0,334	0,112	136,385	0,751	0,751	0,564	306,382
12	0,439	-0,439	0,192	382,119	0,589	-0,589	0,347	513,158	0,527	0,527	0,278	459,502
13	0,204	-0,204	0,042	237,854	0,507	-0,507	0,257	589,635	0,915	0,915	0,837	1065,010
14	0,487	-0,487	0,238	2397,147	0,718	-0,718	0,516	3533,350	0,121	-0,121	0,015	596,286
15	0,073	0,073	0,005	18,471	0,268	-0,268	0,072	68,112	2,551	-2,551	6,509	648,422
16	0,141	0,141	0,020	19,946	0,030	0,030	0,001	4,286	1,186	1,186	1,407	167,202
17	0,003	0,003	0,000	0,802	0,263	-0,263	0,069	82,674	0,812	0,812	0,659	255,374
18	1,087	-1,087	1,182	175,327	1,317	-1,317	1,734	212,346	0,343	0,343	0,118	55,292
19	0,133	-0,133	0,018	102,929	0,450	-0,450	0,203	349,496	0,948	0,948	0,899	735,800
20	0,649	-0,649	0,421	802,473	0,635	-0,635	0,403	784,891	0,060	0,060	0,004	74,493
Média	0,514	0,144	0,456	457,562	0,468	-0,161	0,330	844,681	0,977	0,540	1,321	457,370

Vale ressaltar ainda que, apesar de o $Prompt\ B$ ter apresentado o melhor resultado, esse não poderia ser utilizado em muitos casos, por necessitar de uma estimativa prévia do parâmetro do TRI. Assim, o $Prompt\ C$ se torna uma alternativa promissora por estimar esse parâmetro com base apenas no texto e opções da questão.

5.2. Em que medida essas simulações calibradas com os parâmetros TRI fornecidos pelo INEP conseguem reproduzir a dificuldade observada nas questões reais?

A Tabela 3 apresenta os resultados da conversão para a dificuldade equivalente do parâmetro b_0 das questões do ENEM, em que 1 é fácil, 2 é média e 3 é difícil. Também se apresenta o valor da diferença dessa conversão dos parâmetros b_i (Tabela 1) calculados para diferentes limites para o nível médio de dificuldade. Um valor positivo indica que o nível foi maior que o real e um valor negativo indica que o nível foi menor que o real. A tabela ainda apresenta as métricas de acurácia e erro para esses diferentes cenários.

Os resultados mostram que, em todos os cenários, a taxa de acerto seria maior que um resultado casuístico, dado que todos foram acima do $\frac{1}{3}$ esperado, chegando-se a mais que o dobro para o $Prompt\ B$ com um limite de $[-0,75; 0,75]$. Vale salientar que, apesar de o referido $Prompt\ B$ ter tido os melhores resultados em relação ao erro para a Pergunta de Pesquisa 1, ao considerar o limite padrão de $[-0,50; 0,50]$ para as perguntas de nível médio, esse foi o *prompt* que obteve a pior acurácia.

Tabela 3. Dificuldade calculada a partir de diferentes limites do parâmetro b para a dificuldade média, para os *prompts* A, B e C.

♯	Dificuldade	Limiar de b para dificuldade média.								
		$[-0,5; 0,5]$			$[-0,75; 0,75]$			$[-1,0; 1,0]$		
		ε_A	ε_B	ε_C	ε_A	ε_B	ε_C	ε_A	ε_B	ε_C
1	2	0	+1	0	0	0	0	0	0	0
2	3	-1	0	-1	-1	0	-1	-1	-1	-1
3	3	-1	0	-1	-1	-1	-1	-1	-1	-1
4	3	-1	0	-1	-1	0	-1	-1	-1	-1
5	3	-1	0	0	-1	0	0	-1	0	0
6	3	-1	0	-2	-1	-1	-1	-1	-1	-1
7	3	-1	0	-2	-1	-1	-1	-1	-1	-1
8	3	0	0	0	0	0	0	-1	0	0
9	3	-1	-1	-2	-1	-1	-1	-1	-1	-1
10	2	0	+1	0	0	0	0	0	0	0
11	2	0	+1	-1	0	0	0	0	0	0
12	2	+1	+1	0	0	0	0	0	0	0
13	2	0	+1	-1	0	0	-1	0	0	0
14	2	0	+1	0	0	0	0	0	0	0
15	2	0	+1	+1	0	0	+1	0	0	+1
16	3	0	0	-1	-1	-1	-1	-1	-1	-1
17	2	0	+1	0	0	0	0	0	0	0
18	1	+1	+1	0	+1	+1	0	+1	+1	+1
19	2	0	+1	-1	0	0	-1	0	0	0
20	2	+1	+1	0	0	0	0	0	0	0
	Acurácia	50%	40%	45%	55%	70%	50%	50%	60%	55%
	MAE	0,50	0,60	0,70	0,45	0,30	0,50	0,50	0,40	0,45
	MBE	-0,20	0,50	-0,60	-0,35	-0,20	-0,40	-0,40	-0,30	-0,25
	MSE	0,50	0,60	1,00	0,45	0,30	0,50	0,50	0,40	0,45

Exceto por alguns erros apresentados pelo $Prompt\ C$ no primeiro limiar, todos os demais foram de no máximo um nível para cima ou para baixo, algo que poderia facilmente ocorrer mesmo com a avaliação da dificuldade por profissionais humanos.

Dadas as limitações já citadas do $Prompt\ B$ e a promissora capacidade de análise textual, com inferência dos parâmetros do TRI que o $Prompt\ C$ possui, essa estratégia se mostra promissora para a estimativa do nível de dificuldade. Ressalta-se ainda que a estratégia de ajustar os limites de dificuldade para o nível intermediário pode auxiliar em um ajuste fino do modelo final. Por fim, considera-se que essa estratégia permitiu uma classificação satisfatória do nível de dificuldade das questões do ENEM analisadas, ainda que o tamanho da amostra limite a extrapolação dos resultados.

6. Considerações Finais

Este trabalho apresentou uma estratégia que combina a capacidade de análise textual dos Grandes Modelos de Linguagem (LLMs) com técnicas de simulação para estimar o parâmetro de dificuldade b da Teoria de Resposta ao Item (TRI), visando determinar a dificuldade de questões do ENEM. Os resultados indicam que, embora pequenas discrepâncias tenham sido observadas nas métricas de erro, essas não comprometeram substancialmente a classificação das questões em níveis de dificuldade (fácil, médio e difícil). Além disso, foi possível utilizar um LLM para estimar a dificuldade de uma questão apenas a partir do seu texto e das alternativas, simulando respostas de estudantes com diferentes níveis de habilidade.

Para trabalhos futuros, recomenda-se ampliar a análise para um maior número de questões e áreas de conhecimento, além de explorar outros LLMs, técnicas de *fine-tuning*, *soft prompts* e abordagens como *Retrieval-Augmented Generation* (RAG), com o objetivo de refinar ainda mais a estimativa do parâmetro b e a avaliação automatizada da dificuldade de itens. Essa abordagem mostra-se promissora para apoiar a criação de avaliações escaláveis, consistentes e adaptadas ao perfil dos estudantes.

Referências

Andrade, D. F. d., Tavares, H. R., and Valle, R. C. (2000). *Teoria da Resposta ao Item: Conceitos e Aplicações*. Associação Brasileira de Estatística, São Paulo.

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

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.

Bussab, W. d. O. and Morettin, P. A. (2017). *Estatística básica*. Saraiva Educação, São Paulo, 9 edition.

Chen, C. H. and Shiu, M. F. (2025). Kaqg: A knowledge-graph-enhanced rag for difficulty-controlled question generation. *arXiv preprint arXiv:2505.07618*.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc, Newbury Park, CA.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition. URL: `https://otexts.com/fpp2/`.

Jain, Y., Hollander, J., He, A., Tang, S., Zhang, L., and Sabatini, J. (2025). Exploring the potential of large language models for estimating the reading comprehension question difficulty. *arXiv.org*.

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*.

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.

Marinho, W., Clua, E. W., Martí, L., and Marinho, K. (2023). Predicting item response theory parameters using question statements texts. *International Conference on Learning Analytics and Knowledge*.

Mulla, N. and Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*.

Ogbonna, J. and Opara, I. (2018). Estimating standard errors of irtparameters of mathematics achievement test using three parameter model. *IOSR Journal of Research & Method in Education (IOSR-JRME)*, 8(2):01–07.

Tomikawa, Y. and Uto, M. (2024). Difficulty-controllable multiple-choice question generation for reading comprehension using item response theory. In *AIED Companion*.