Estimando Parâmetros da Teoria de Resposta ao Item com NLP e Aprendizado de Máquina

William Oliveira da Costa e Silva[1], **Filipe Dwan Pereira**[1], **Rafael Mello**[2]

[1]Universidade Federal de Roraima (UFRR)

[2]CESAR

{filipe.dwan,rafael.mello}@cesar.edu.br

{william,filipe.dwan}@ufrr.br

Abstract. *Item calibration in Item Response Theory (IRT) is a costly process that depends on student responses. We propose a methodology to predict the difficulty (b) and discrimination (a) parameters of new items from their text, eliminating this dependency. The method trains a regression model using real IRT parameters as a target and a feature set that combines rubrics generated by a Large Language Model (LLM) with embeddings. The results indicate that the interpretable rubrics are as predictive as the embeddings, validating a workflow for a priori item calibration that can streamline the creation of assessments.*

Resumo. *A calibração de itens da Teoria de Resposta ao Item (TRI) é um processo custoso que depende de respostas de estudantes. Propomos uma metodologia para predizer os parâmetros de dificuldade (b) e discriminação (a) de itens novos a partir do texto, eliminando essa dependência. O método treina um modelo de regressão usando parâmetros reais da TRI como alvo e um conjunto de features que une rubricas geradas por um Modelo de Linguagem Amplo (LLM) com embeddings. Os resultados indicam que as rubricas interpretáveis são preditivas bem como os embeddings, validando um fluxo de trabalho para calibração a priori de itens que pode agilizar a criação de avaliações.*

1. Introdução

A Teoria de Resposta ao Item (TRI) é um paradigma psicométrico fundamental para a modelagem da relação entre o nível de habilidade latente de um indivíduo e a probabilidade de ele responder corretamente a um item de um teste [Lord 2012]. A calibração de itens, processo de estimação dos parâmetros dos itens (e.g., dificuldade, discriminação, acerto casual), é uma etapa no desenvolvimento de avaliações educacionais de larga escala e bancos de itens. Tradicionalmente, a estimação desses parâmetros requer a aplicação dos itens a uma amostra considerável de respondentes, o que pode ser um processo custoso e demorado, especialmente para itens novos ou em desenvolvimento [Yancey et al. 2024].

Modelos de Linguagem Grandes (LLMs - do inglês *Large Language Models*) e técnicas de representação textual, como os *embeddings*, têm aprimorado o Processamento de Linguagem Natural (PLN), oferecendo capacidades na compreensão e geração de texto [Rodrigues et al. 2024b, Mello et al. 2024]. No contexto educacional, tanto os LLMs quanto os *embeddings* têm sido explorados para diversas tarefas. Os

DOI: 10.5753/sbie.2025.12767

LLMs, por exemplo, são utilizados na geração de questões e na avaliação de respostas [Liu et al. 2025, Rodrigues et al. 2024a, Gurdil et al. 2024]. Paralelamente, os *embeddings* fornecem representações vetoriais densas que capturam nuances semânticas do texto dos itens. A combinação do poder de extração de características contextuais dos LLMs com a representacional dos *embeddings* abre novas e promissoras possibilidades para a estimação da dificuldade das questões (Question Difficulty Estimation from Text - QDET) [Benedetto et al. 2023, Yancey et al. 2024], permitindo uma análise profunda e multifacetada das propriedades psicométricas inferidas a partir do texto.

Este trabalho propõe uma nova metodologia para predizer os parâmetros da TRI para itens de múltipla escolha, utilizando uma combinação de TRI tradicional, LLMs e modelos preditivos de aprendizado de máquina. A principal contribuição deste artigo reside na introdução de uma abordagem híbrida que integra (i) a robustez da TRI na calibração inicial de itens, (ii) o poder dos LLMs para engenharia de características textuais baseada em rubricas semânticas, e (iii) *embeddings* para capturar representações latentes do conteúdo dos itens. Demonstramos que esta combinação permite a predição dos parâmetros da TRI para novos itens, superando uma lacuna na literatura referente à necessidade de grandes amostras de respondentes para calibração. Ao contrário de métodos que se baseiam unicamente em características textuais superficiais ou *embeddings* como caixa-preta, nossa abordagem visa incorporar um nível de granularidade e interpretabilidade através das rubricas geradas pelo LLM, ao mesmo tempo que se beneficia da riqueza representacional dos *embeddings*. Esta metodologia tem o potencial de acelerar o processo de desenvolvimento e atualização de bancos de itens, tornando a avaliação educacional mais ágil e adaptável.

A abordagem inicia com a estimação dos parâmetros da TRI (especificamente, dificuldade 'b' e discriminação 'a' sob o modelo 2PL - detalhado na seção 2) para um conjunto inicial de itens, utilizando respostas de alunos a provas do Exame Nacional de Desempenho dos Estudantes (Enade) na área de engenharia. Estes parâmetros estimados servem como *ground truth* para a fase subsequente. Em seguida, um LLM é empregado para analisar o texto dos enunciados e alternativas, extraindo pontuações para um conjunto de rubricas predefinidas que hipoteticamente se relacionam com a dificuldade do item (e.g., nível na taxonomia de Bloom, necessidade de análise gráfica, etc.). Adicionalmente, são gerados *embeddings* a partir do texto dos itens e suas alternativas. Estes *embeddings* são então reduzidos em dimensionalidade via Análise de Componentes Principais (PCA). Finalmente, as pontuações das rubricas geradas pelo LLM e os *embeddings* de baixa dimensão são utilizados como variáveis preditoras em um modelo de regressão para estimar os parâmetros da TRI (dificuldade e discriminação) de novos itens, para os quais não existem dados de resposta prévios.

2. Referencial Teórico

2.1. Teoria de Resposta ao Item

A TRI engloba uma família de modelos matemáticos que buscam descrever a relação entre traços latentes (e.g., habilidades, atitudes) de indivíduos e suas respostas a itens de um instrumento de medida [Lord 2012]. Diferentemente da Teoria Clássica dos Testes (TCT), que foca em escores totais, a TRI modela a probabilidade de uma resposta específica a um item em função dos parâmetros do item e da habilidade do respondente.

Para itens dicotômicos (certo/errado), modelos comuns incluem o modelo logístico de um parâmetro (1PL ou modelo de Rasch), que considera apenas a dificuldade do item (b_i); o modelo logístico de dois parâmetros (2PL), que adiciona o parâmetro de discriminação do item (a_i); e o modelo logístico de três parâmetros (3PL), que inclui também o parâmetro de acerto casual (c_i). A probabilidade de um indivíduo j com habilidade θ_j acertar um item i em um modelo 2PL é dada por:

$$P(X_{ij} = 1|\theta_j, a_i, b_i) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

onde $X_{ij} = 1$ indica uma resposta correta. O parâmetro b_i representa a dificuldade do item (o nível de habilidade onde a probabilidade de acerto é 0.5), e a_i indica o quão bem o item discrimina entre indivíduos com diferentes níveis de habilidade. Neste trabalho, utilizamos o modelo 2PL para estimar os parâmetros dos itens do Enade, uma vez que as questões são de múltipla escolha e o parâmetro de acerto casual (c) pode ser, em princípio, estimado ou fixado. A estimação desses parâmetros é realizada usando abordagens Bayesianas, com a inferência variacional implementada na biblioteca `py-irt` [Natesan et al. 2016].

2.2. Estimação da Dificuldade da Questão a partir do Texto (QDET)

A Estimação da Dificuldade da Questão a partir do Texto (Question Difficulty Estimation from Text - QDET) é uma área de pesquisa que aplica técnicas de PLN para prever a dificuldade de questões educacionais analisando seu conteúdo textual [Benedetto et al. 2023]. O objetivo é superar as limitações dos métodos tradicionais de calibração de itens, que dependem da coleta de respostas de um grande número de examinados. As abordagens de QDET variam amplamente, utilizando desde características linguísticas tradicionais (e.g., contagem de palavras, complexidade sintática, legibilidade) até representações mais sofisticadas baseadas em *embeddings* de palavras ou sentenças e modelos de aprendizado profundo [Benedetto 2023]. Os *embeddings*, em particular, são vetores densos que capturam o significado semântico do texto, permitindo que modelos de aprendizado de máquina identifiquem padrões relacionados à representação textual.

3. Trabalhos Relacionados

A tarefa de estimar parâmetros de itens da TRI sem a necessidade de respostas humanas extensivas, tem sido um foco crescente de pesquisa [Byrd and Srivastava 2022, Benedetto et al. 2023, Benedetto 2023, Yancey et al. 2024], impulsionada pelos avanços em PLN e aprendizado de máquina.

[Benedetto et al. 2023] apresenta uma revisão de abordagens para QDET, categorizando-as com base nas características das questões e nas técnicas de PLN empregadas. Trabalhos frequentemente utilizavam características linguísticas superficiais e modelos de regressão lineares. Mais recentemente, modelos baseados em *embeddings* e redes neurais profundas têm mostrado desempenho superior [Benedetto 2023]. Por exemplo, [Byrd and Srivastava 2022] exploraram modelos para estimar diretamente a dificuldade e a discriminação de questões de múltipla escolha, correlacionando esses parâmetros com características das perguntas, respostas e contextos associados. Nossa abordagem se diferencia ao incorporar explicitamente rubricas semânticas extraídas por LLMs como

um conjunto de características, além dos *embeddings*, e ao aplicar este método no contexto específico do Enade para engenharia, usando parâmetros da TRI obtidos via TRI tradicional como *ground truth*.

[Yancey et al. 2024] propuseram o BERT-IRT, um modelo explanatório de TRI que utiliza *embeddings* do BERT e características de PLN para acelerar a pilotagem de itens, demonstrando como essa abordagem pode reduzir a necessidade de longos pilotos sem sacrificar a validade. Enquanto o BERT-IRT foca em enriquecer o modelo IRT com informações textuais para uma melhor estimação com dados de resposta, nosso trabalho foca na predição dos parâmetros da TRI para itens novos (sem dados de resposta), utilizando os parâmetros da TRI como alvos de um modelo de aprendizado de máquina treinado com características textuais, incluindo anotações de LLMs. [Lalor et al. 2019] investigaram a aprendizagem de parâmetros latentes da TRI sem padrões de resposta humana, utilizando *multidões artificiais* (modelos DNN) para gerar dados de resposta. Embora o objetivo de evitar a dependência de dados humanos seja similar, nossa metodologia difere ao focar na extração de características do texto do item para predição direta d -s parâmetros, em vez de simular respondentes.

Por fim, o uso de LLMs na avaliação educacional e psicometria está se expandindo rapidamente. [Liu et al. 2025] realizaram uma análise psicométrica do uso de LLMs como respondentes para avaliação de itens, investigando se LLMs podem gerar respostas com propriedades psicométricas comparáveis às de humanos. Eles descobriram que alguns LLMs exibem proficiência similar ou superior à de estudantes universitários em álgebra, mas com distribuições de proficiência mais estreitas. [Gurdil et al. 2024] exploraram a eficácia de LLM na geração de dados no escopo da TRI, concluindo que LLMs podem ser ferramentas úteis no desenvolvimento de algoritmos de geração de dados, embora com algumas limitações na replicação precisa das condições de simulação dos parâmetros dos itens. Estes trabalhos destacam o potencial dos LLMs em interagir com construtos psicométricos. Nossa pesquisa contribui para esta linha ao utilizar LLMs não como respondentes ou geradores de dados de resposta, mas como ferramentas de engenharia de características, extraindo metadados semânticos (rubricas) dos itens que são posteriormente usados para prever seus parâmetros psicométricos.

4. Metodologia

A metodologia proposta para a predição de parâmetros da TRI para novas questões é composta por três etapas principais, conforme detalhado nesta seção. Primeiramente, realizamos a estimação dos parâmetros da TRI para um conjunto existente de itens utilizando dados de resposta de estudantes. Em seguida, procedemos com a engenharia de características, que envolve a extração de metadados textuais através de um LLM e a geração de *embeddings* a partir do conteúdo das questões. Finalmente, um modelo de aprendizado de máquina é treinado para predizer os parâmetros da TRI de itens inéditos com base nessas características.

4.1. Conjunto de Dados

O conjunto de dados[1] utilizado neste estudo é derivado do Enade, uma avaliação em larga escala aplicada a estudantes concluintes de cursos de graduação no Brasil. Optou-se por

[1] www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade

este conjunto por se tratar de dados públicos que abrangem estudantes de diversas regiões, instituições de ensino e condições socioeconômicas. Ademais, as provas costumam ocorrer em ambientes supervisionados, o que minimiza distorções nos cálculos dos parâmetros da TRI decorrentes de consultas indevidas durante a resolução do exame.

Em relação aos cursos, optou-se por escolher respondentes relacionados aos cursos de engenharia. Tal fato se justifica pela complexidade inerente aos enunciados das questões destes cursos, que frequentemente combinam elementos textuais com a interpretação de recursos visuais como gráficos, diagramas e fluxogramas. Além disso, é uma área do conhecimento pouco explorada por trabalhos semelhantes, os quais concentram sua atenção em questões relacionadas ao aprendizado de idiomas, ciência da computação e medicina [AlKhuzaey et al. 2024].

O processo de tratamento dos dados brutos envolveu:

1. Carregamento dos microdados do Enade referentes aos anos de 2023, 2019, 2017, 2014 e 2011;
2. Filtragem para selecionar exclusivamente os respondentes e itens pertencentes aos cursos de engenharia. Essa seleção foca em um único domínio para ter uma maior homogeneidade do construto latente avaliado, algo importante para modelos de TRI.
3. Filtragem para selecionar apenas respondentes com participação válida;
4. Construção de uma matriz de respostas esparsa, em que cada coluna representa a resposta de um estudante para uma questão, com valores indicando resposta incorreta (0), correta (1) ou anulada (8 ou 9). Questões que foram anuladas (código 8 ou 9) foram removidas para garantir a qualidade da estimação dos parâmetros da TRI;
5. Geração de um arquivo no formato `.jsonlines` contendo a matriz de respostas.

Este conjunto de dados de respostas serve como base para a Etapa 1 (seção 4.2) da nossa metodologia.

4.2. Etapa 1: Estimação dos Parâmetros da TRI (Geração do *Ground Truth*)

Nesta etapa, os parâmetros dos itens da TRI são estimados a partir do arquivo (`.jsonlines`) que contém a matriz de respostas dos estudantes. Estes parâmetros, notadamente a dificuldade (b) e a discriminação (a), servem como *ground truth* para o treinamento do modelo preditivo subsequente.

A matriz de respostas, gerada conforme descrito na Seção 4.1, é o insumo principal. Cada entrada (j, i) na matriz representa a resposta do estudante j ao item i.

Conforme apontado na seção 2.1,para a estimação dos parâmetros, adotou-se o modelo 2PL da TRI, adequado para itens de múltipla escolha onde o acerto casual não é explicitamente modelado como um terceiro parâmetro (c) neste estágio, visando simplificação e robustez devido ao tamanho do conjunto de dados.

4.3. Etapa 2: Engenharia de Características para Predição

Com os parâmetros da TRI estimados, a próxima etapa consiste em gerar um conjunto de características (features) a partir do texto das questões (enunciado e alternativas) que possam ser preditivas desses parâmetros. Esta etapa visa capturar as propriedades textuais que influenciam a dificuldade e a discriminação de um item.

Inicialmente, realiza-se a extração do texto dos enunciados e das alternativas. Devido à quantidade de questões analisadas, a extração manual dos textos se mostrou inviável. Para superar essa limitação, optou-se pela utilização de LLM multimodal para extração automática dos textos e, se houver, descrição textual detalhada das imagens, diagramas, gráficos e fluxogramas com base no arquivo `.pdf` das provas. Para a escolha do modelo, optou-se pela análise de uma combinação do desempenho no *benchmark* [LMArena 2025] e custo por milhão de tokens. O modelo que se mostrou mais adequado foi o `gemini-2.5-flash-preview-05-20` [LMArena 2025].

Após a extração do texto dos enunciados e alternativas, e a obtenção da descrição textual detalhada dos recursos visuais, procedeu-se à geração dos embeddings. O modelo utilizado para essa etapa foi o `text-embedding-004` da API do Gemini. A escolha deste modelo é justificada por seu desempenho de ponta em *benchmarks* de representação semântica, como o *Massive Text Embedding Benchmark* (MTEB).

Após a geração do embedding do enunciado, que é fornecido a um modelo de regressão, analisou-se a similaridade de cosseno entre todos os pares de questão para identificar possíveis questões duplicadas no conjunto de dados. Ademais, observou-se que o embedding final possui 768 dimensões. No entanto, com o intuito de evitar a "*maldição da dimensionalidade*"descrita em [Köppen 2000], optou-se por aplicar a PCA para redução de sua dimensionalidade para 80 parâmetros.

Quanto aos embeddings das alternativas, o objetivo principal é gerar features com alta capacidade preditiva, tais como: a similaridade semântica média entre as alternativas, e as similaridades semânticas máxima, mínima e média entre a alternativa correta e os distratores. Essas escolhas foram motivadas por resultados promissores encontrados em trabalhos como [Hsu et al. 2018] e [Susanti et al. 2017].

Em seguida, empregou-se o modelo (`gemini-2.5-flash-preview-05-20`) para atribuição de metadados aos itens com base em um conjunto de rubricas predefinidas. Essas rubricas são projetadas para quantificar dimensões textuais e cognitivas hipoteticamente relacionadas à complexidade dos itens, as quais são: tipo de raciocínio envolvido (taxonomia de Bloom), necessidade de análise de elementos gráficos para resolução de questão, número de etapas cognitivas envolvidas, presença de conhecimento interdisciplinar e necessidade de manipulações algébricas complexas. O LLM é instruído a pontuar cada item (enunciado e alternativa) em relação a cada uma dessas rubricas.

Para melhorar a acurácia da LLM na extração dos metadados, utilizou-se uma técnica descrita em [Wei et al. 2022] conhecida como *chain-of-thought*. No caso, além da tarefa solicitada, fornece-se, no *prompt*, uma descrição - em linguagem natural - das etapas que a LLM deve seguir para chegar ao resultado. Os prompts e códigos utilizados serão disponibilizados após a aprovação do artigo.

4.4. Etapa 3: Modelo Preditivo dos Parâmetros da TRI

A etapa final envolve a construção e o treinamento de um modelo de aprendizado de máquina para predizer os parâmetros da TRI (a e b) para itens novos, utilizando as características geradas na Etapa 2.

Para cada item do conjunto original (para o qual os parâmetros da TRI foram estimados na Etapa 1), o vetor de características é formado pela concatenação das pontuações

das rubricas obtidas pelo LLM e os *embeddings* de baixa dimensão (pós-PCA). Os parâmetros da TRI estimados (a_i e b_i) para cada item são utilizados como as variáveis alvo (rótulos). O conjunto de dados é então dividido em subconjuntos de treinamento e teste com uma proporção de 80% para treino e 20% para teste.

Dois modelos de regressão separados são treinados: um para predizer o parâmetro de dificuldade (b) e outro para predizer o parâmetro de discriminação (a). Como o conjunto de dados é pequeno, optou-se pela utilização de um modelo `RandomForestRegressor` do `scikit-learn` para esta tarefa. A escolha decorreu da alta precisão, robustez a ruídos que esses modelos possuem [Liu et al. 2012]. O modelo aprende a mapear as características textuais (rubricas do LLM e *embeddings* PCA) para os respectivos parâmetros da TRI.

$$\hat{b}_i = f_{regressor_b}(Rubricas_i, Embeddings_{PCA_i})$$

$$\hat{a}_i = f_{regressor_a}(Rubricas_i, Embeddings_{PCA_i})$$

onde $\hat{b}_i$ e $\hat{a}_i$ são as predições para a dificuldade e discriminação do item i, respectivamente.

A performance dos modelos preditivos é avaliada no conjunto de teste utilizando métricas padrão de regressão: Erro Quadrático Médio (Mean Squared Error - MSE), Erro Absoluto Médio (Mean Absolute Error - MAE) e o Coeficiente de Determinação (R^2):

$$MSE = \frac{1}{N_{teste}} \sum_{i=1}^{N_{teste}} (y_i - \hat{y}_i)^2$$

$$MAE = \frac{1}{N_{teste}} \sum_{i=1}^{N_{teste}} |y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{teste}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{teste}} (y_i - \bar{y})^2}$$

onde y_i é o valor real do parâmetro TRI (dificuldade ou discriminação) para o item i no conjunto de teste, $\hat{y}_i$ é o valor predito pelo modelo, e $\bar{y}$ é a média dos valores reais no conjunto de teste. Estas métricas quantificam a precisão das predições dos parâmetros para itens não vistos durante o treinamento.

Por fim, para buscar minimizar os erros das predições, a técnica `GridSearchCV` foi utilizada para procurar qual combinação de hiperparâmetros geraria os melhores resultados preditivos para os parâmetros a e b. As combinações foram realizadas para os hiperparâmetros dispostos na Tabela 1.

5. Resultados

Esta seção apresenta os resultados obtidos nas etapas de análise do conjunto de dados e de modelagem preditiva dos parâmetros da TRI. Inicialmente, são detalhadas as características do conjunto de questões processado. Em seguida, são expostos os resultados do modelo de aprendizado de máquina para a predição dos parâmetros de discriminação (a) e dificuldade (b), incluindo as métricas de desempenho e as configurações otimizadas de hiperparâmetros.

Tabela 1. Hiperparâmetros Utilizados no Grid Search para Otimização do RandomForestRegressor.

Hiperparâmetro	Valores Testados	Descrição
`n_estimators`	{50, 100, 200, 300, 400, 500}	Número de árvores na floresta.
`max_depth`	{5, 10, 15, 20, 25, None}	Profundidade máxima de cada árvore. (None indica profundidade ilimitada).
`min_samples_leaf`	{1, 5, 10}	Número mínimo de amostras necessárias para estar em um nó folha.
`max_features`	{1.0, 'sqrt', 'log2'}	Número de características a considerar em cada split.

5.1. Características do conjunto de dados

Após o processamento das questões do ENADE para os cursos de engenharia e para os anos de 2023, 2019, 2017 e 2014, conforme descrito nas Seções 4.2 e 4.3, um total de 811 questões únicas foi obtido. Do total, 11,8% foram classificadas como *complexa* e 88,2% como *não complexas*. Em relação à distribuição de nível na taxonomia de Bloom, 30,8% foram classificadas como *Avaliar*, 28,2% como *Aplicar*, 19,1% como *Analisar*, 11,7% como *Compreender*, 9,1% como *Memorizar* e 1,0% como *Criar*. Quanto à necessidade de análise gráfica, 54,9% *não necessita* e 45,% *necessita*. No mais, 82,9% das questões foram classificadas como *não interdisciplinares* e 17,1% como *interdisciplinares*. Finalmente, a Figura 1 apresenta o histograma do número de etapas cognitivas necessárias para a resolução das questões, com uma maior concentração em valores menores do 10.

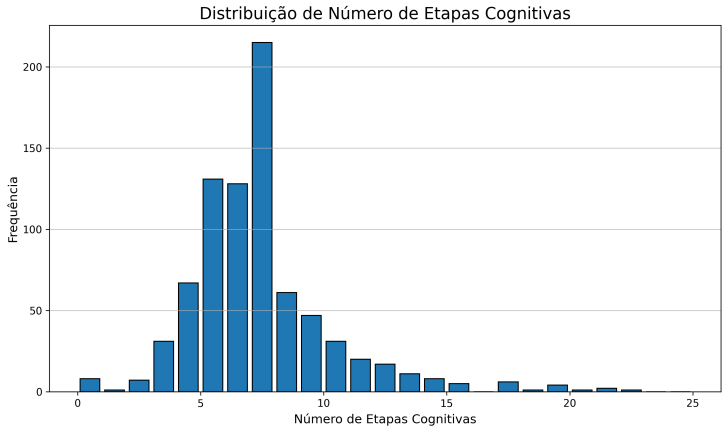

Figura 1. Histograma - etapas cognitivas

5.2. Resultados do modelo preditivo para o parâmetro de discriminação a

Após a aplicação do `GridSearchCV` para a combinação de diferentes hiperparâmetros conforme descrito na Tabela 1, os melhores parâmetros encontrados para predição do parâmetro de discriminação a da TRI são descritos na Tabela 2.

Tabela 2. Melhores Hiperparâmetros Encontrados para o RandomForestRegressor na Predição do Parâmetro de Discriminação (a)

Hiperparâmetro	Melhor valor
n_estimators	500
max_depth	25
min_samples_leaf	5
max_features	1.0

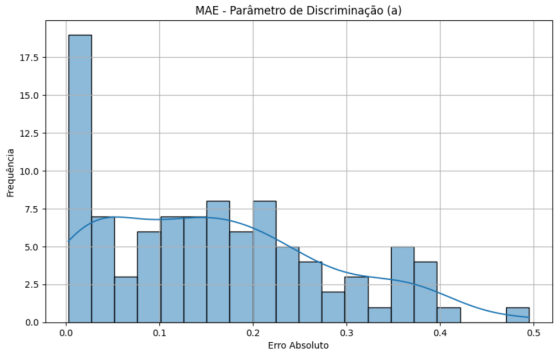

Figura 2. Distribuição residual de erros - Parâmetro de Discriminação (a)

Os resultados encontrados para MAE, MSE e R^2 são descritos na Tabela 3. A distribuição residual dos erros é descrita na Figura 2.

Tabela 3. Métricas para o Parâmetro de Discriminação (a)

Métricas	Resultados
MAE	0.1604
MSE	0.0397
R^2	0.1243

5.3. Resultados do modelo preditivo para o parâmetro de dificuldade b

Após a aplicação do GridSearchCV para a combinação de diferentes hiperparâmetros conforme descrito na Tabela 1, os melhores parâmetros encontrados para predição do parâmetro de dificuldade b da TRI são descritos na Tabela 4.

Tabela 4. Melhores Hiperparâmetros Encontrados para o RandomForestRegressor na Predição do Parâmetro de Dificuldade (b)

Hiperparâmetro	Melhor valor
n_estimators	500
max_depth	15
min_samples_leaf	5
max_features	1.0

Os resultados encontrados para MAE, MSE e R^2 são descritos na Tabela 5. A distribuição residual dos erros é descrita na Figura 3.

Tabela 5. Métricas para o Parâmetro de Dificuldade (b)

Métricas	Resultados
MAE	0.6783
MSE	0.6761
R^2	0.0519

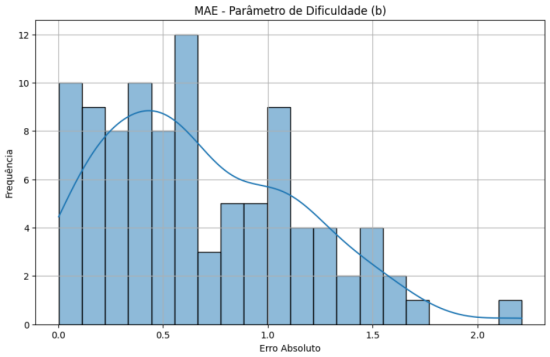

Figura 3. Distribuição residual de erros - Parâmetro de Dificuldade (b)

5.4. Importância das features

A Tabela 6 demonstra as 15 features mais importantes para a predição do parâmetro de discriminação (a) e dificuldade (b). Em que `pc` referem-se aos componentes principais extraídos dos embeddings dos enunciados após aplicação da técnica PCA.

Tabela 6. Ranqueamento das 15 *features* mais importantes para os parâmetros de discriminação (a) e dificuldade (b)

Posição	Feature para (a)	Feature para (b)
1	pc1	pc1
2	COMPLEXIDADE ALGÉBRICA	COMPLEXIDADE ALGÉBRICA
3	pc44	pc57
4	pc24	pc26
5	pc48	pc19
6	pc3	pc6
7	pc23	pc22
8	pc32	pc5
9	pc4	pc39
10	pc33	pc34
11	pc17	pc4
12	pc16	pc9
13	pc71	pc63
14	TAXONOMIA DE BLOOM	pc13
15	pc39	pc64

6. Discussões

6.1. Interpretação dos Resultados Preditivos

O ponto mais significativo extraído da análise de importância das características (Tabela 6) é a sinergia entre as representações latentes (componentes principais dos *embeddings*) e os meta-dados semânticos gerados pelo LLM. A característica `COMPLEXIDADE ALGÉBRICA`, uma rubrica interpretável, posicionou-se como a segunda mais importante para a predição de ambos os parâmetros, discriminação (a) e dificuldade (b). Este resultado demonstra que características explícitas e conceitualmente ricas, extraídas por um LLM, podem rivalizar e complementar a informação contida nas representações densas e "caixa-preta"dos *embeddings*. Enquanto os componentes principais `pc1` dominam o ranking, a forte presença de uma rubrica semântica sugere que a combinação de ambas as abordagens captura um espectro mais amplo de sinais preditivos do que cada uma isoladamente. A inclusão da `TAXONOMIA DE BLOOM` entre as 15 características mais relevantes para o parâmetro a reforça ainda mais essa conclusão.

Ao analisar o desempenho preditivo para o parâmetro de discriminação (a), o modelo alcançou um desempenho modesto, conforme indicado pelo coeficiente de determinação ($R^2 = 0.1243$) na Tabela 3. Embora este valor sugira que o modelo explica apenas uma pequena porção da variância total, a análise da distribuição dos erros residuais (Figura 2) fornece um contraponto otimista. A distribuição é acentuadamente centrada em zero e aproximadamente simétrica, indicando uma ausência de viés sistemático nas predições. Ou seja, o modelo não tende a superestimar ou subestimar consistentemente o parâmetro a.

A predição do parâmetro de dificuldade (b) provou ser uma tarefa ainda mais desafiadora, com um R^2 de 0.0519 (Tabela 5). Este resultado está alinhado com a literatura, que frequentemente aponta a dificuldade como um construto psicométrico complexo e de difícil modelagem apenas a partir do texto [Benedetto et al. 2023]. No entanto, similarmente à análise do parâmetro a, a Figura 3 revela que, apesar da baixa explicação da variância, uma parcela substancial dos erros residuais se concentra em um intervalo relativamente estreito em torno de zero (com a maioria sendo menor que 0.5 em valor absoluto). Isso sugere que, embora o modelo possa cometer erros maiores em alguns itens, para uma porção significativa do conjunto de teste, as predições estão razoavelmente próximas dos valores reais (estimados diretamente a partir das respostas dos alunos).

6.2. Principais Contribuições

Este trabalho oferece várias contribuições para a intersecção entre psicometria e inteligência artificial:

- Metodologia Híbrida: Propomos e validamos um fluxo de trabalho que integra de forma sinérgica a estimação de parâmetros da TRI tradicional (usando `py-irt`), a engenharia de características interpretáveis via LLMs (rubricas) e a representação de texto via *embeddings*. Esta abordagem multifacetada constitui um avanço em relação a métodos que dependem de uma única fonte de características.
- Validação da Sinergia entre LLM e Embeddings: Fornecemos evidências empíricas de que meta-dados semânticos gerados por LLMs não são redundantes em relação aos *embeddings*. Pelo contrário, eles adicionam um valor preditivo significativo e interpretável, como demonstrado pelo alto ranqueamento da

COMPLEXIDADE ALGÉBRICA. Isso abre caminho para modelos mais robustos e menos ”caixa-preta”.

- Predição de Parâmetros de Itens em um Cenário de *Cold Start*: A contribuição central deste artigo é a demonstração de um método para predizer os parâmetros de discriminação (a) e dificuldade (b) de itens completamente novos, analisando exclusivamente seu conteúdo textual. Nossa abordagem elimina a necessidade de coletar dados de resposta de estudantes para a calibração inicial de novos itens, resolvendo um dos principais gargalos logísticos e financeiros no desenvolvimento de avaliações. Isso permite, por exemplo, que um banco de questões seja expandido com itens cujas propriedades psicométricas são estimadas a priori, antes de sua aplicação efetiva.

Observe que a capacidade de estimar preliminarmente os parâmetros de novos itens sem a necessidade de uma rodada de pré-testagem com respondentes humanos pode agilizar o processo de desenvolvimento e manutenção de bancos de itens. Instituições de ensino e empresas de avaliação podem utilizar este método para uma triagem automática de itens, identificando aqueles que provavelmente terão parâmetros psicométricos desejáveis (e.g., dificuldade média, alta discriminação) antes de investi-los em etapas de validação mais custosas.

7. Conclusões e Trabalhos Futuros

Este trabalho propôs e avaliou uma metodologia para a predição dos parâmetros de discriminação (a) e dificuldade (b) da TRI com base no enunciado de questões. A abordagem integrou o uso de LLMs para a engenharia de características, incluindo metadados e *embeddings* densos, com modelos preditivos de aprendizado de máquina. A análise apresentada demonstrou a viabilidade da metodologia proposta e o potencial das características extraídas por LLMs para inferir traços psicométricos. Conclui-se que a combinação de features interpretáveis (como Complexidade Algébrica e Taxonomia de Bloom) com representações densas (componentes de PCA) é uma direção promissora para a automatização da estimação de parâmetros da TRI, fornecendo insights valiosos sobre a natureza da dificuldade e discriminação de itens.

Com base nos resultados e limitações identificadas, trabalhos futuros podem seguir diversas direções. Uma iniciativa crucial seria a expansão e diversificação do conjunto de dados, com a criação de *datasets* públicos de engenharia mais amplos para que o modelo aprenda com um maior volume de dados e um espectro mais vasto de complexidades. Adicionalmente, a otimização da engenharia de características, por meio do *fine-tuning* de LLMs para a tarefa de extração de rubricas, poderia aumentar a acurácia e a granularidade dos metadados. Outra área promissora é a investigação de métodos para a representação multimodal direta, que permitiria a incorporação de informações visuais (imagens, gráficos) sem depender de descrições textuais, evitando assim a perda de nuances. Além disso, uma análise de erros mais aprofundada, idealmente em colaboração com especialistas em psicometria, poderia identificar padrões em categorias de itens onde a predição é mais desafiadora. Finalmente, a metodologia poderia ser estendida para incluir a predição do parâmetro de acerto casual (c) do modelo 3PL, o que é particularmente relevante para questões de múltipla escolha.

Referências

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.

Benedetto, L. (2023). A quantitative study of nlp approaches to question difficulty estimation. In *International Conference on Artificial Intelligence in Education*, pages 428–434. Springer.

Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., and Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Byrd, M. and Srivastava, S. (2022). Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130.

Gurdil, H., Soguksu, Y. B., Salihoglu, S., and Coskun, F. (2024). Integration of artificial intelligence in educational measurement: Efficacy of chatgpt in data generation within the scope of item response theory. *arXiv preprint arXiv:2402.01731*.

Hsu, F.-Y., Lee, H.-M., Chang, T.-H., and Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.

Köppen, M. (2000). The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8.

Lalor, J. P., Wu, H., and Yu, H. (2019). Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, page 4240.

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.

Liu, Y., Wang, Y., and Zhang, J. (2012). New machine learning algorithm: random forest. In *Proceedings of the Third International Conference on Information Computing and Applications*, ICICA'12, page 246–252, Berlin, Heidelberg. Springer-Verlag.

LMArena (2025). Overview Leaderboard — LMArena. Online.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Mello, R. F., Rodrigues, L., Cabral, L., Pereira, F. D., Júnior, C. P., Gasevic, D., and Ramalho, G. (2024). Prompt engineering for automatic short answer grading in brazilian portuguese. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1730–1743. SBC.

Natesan, P., Nandakumar, R., Minka, T., and Rubright, J. D. (2016). Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., and Mello, R. F. (2024a). Assessing the quality of automatic-generated short answers using gpt-4. *Computers and Education: Artificial Intelligence*, 7:100248.

Rodrigues, L., Pereira, F. D., Cabral, L., Ramalho, G., Gasevic, D., and Mello, R. F. (2024b). Can gpt4 answer educational tests? empirical analysis of answer quality based on question complexity and difficulty. In *International Conference on Artificial Intelligence in Education*, pages 192–205. Springer.

Susanti, Y., Tokunaga, T., Nishikawa, H., and Obari, H. (2017). Controlling item difficulty for automatic vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, 12(1):25.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yancey, K. P., Runge, A., Laflair, G., and Mulcaire, P. (2024). Bert-irt: Accelerating item piloting with bert embeddings and explainable irt models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438.