Predição de Dificuldade de Questões de Ciências Humanas do ENEM

Alexandre E. de Souza[1], **Gustavo S. de Oliveira**[1], **Lucas I. C. Ciziks**[1]

[1]Instituto de Ciências Matemáticas e Computação
Universidade de São Paulo (USP)

{alexandre_souza,gustavo.oliveira03, luciziks}@usp.br

Abstract. *The development of large-scale educational tests, such as the National High School Exam (ENEM), faces significant logistical and financial challenges, notably the pre-testing process required to calibrate item difficulty, which is costly and compromises the reusability of the questions. As an alternative, this work investigates the feasibility of predicting the difficulty parameter (parameter B of Item Response Theory) for questions in the Human Sciences area, based exclusively on their textual characteristics through Natural Language Processing (NLP) techniques. The study aims to reproduce and expand upon the methodology proposed by [Jaloto et al. 2023], which demonstrated the potential of this approach. The main objective is to develop a robust predictive model capable of estimating an item's difficulty before its empirical application, which would allow for more precise control over the test's difficulty level, resource optimization, and an increase in exam security.*

Resumo. *A elaboração de testes educacionais em larga escala, como o Exame Nacional do Ensino Médio (ENEM), enfrenta desafios logísticos e financeiros significativos, notadamente o processo de pré-testagem para calibrar a dificuldade dos itens, que é oneroso e compromete a reutilização das questões. Como alternativa, este trabalho investiga a viabilidade de prever o parâmetro de dificuldade (parâmetro B da Teoria de Resposta ao Item) de questões da área de Ciências Humanas, baseando-se exclusivamente em suas características textuais por meio de técnicas de Processamento de Linguagem Natural (PLN). O estudo busca reproduzir e expandir a metodologia proposta por [Jaloto et al. 2023], que demonstrou a potencialidade da abordagem. O objetivo central é desenvolver um modelo preditivo robusto capaz de estimar a dificuldade de um item antes de sua aplicação empírica, o que permitiria um controle mais preciso sobre o nível da prova, a otimização de recursos e um aumento na segurança do exame.*

1. Introdução

A previsão da dificuldade de questões em exames padronizados, como o ENEM, tem despertado interesse na interseção entre estatística e processamento de linguagem natural. Estudos recentes têm explorado a viabilidade de estimar o parâmetro de dificuldade (b), definido pela Teoria de Resposta ao Item (TRI), com base apenas nas características textuais dos enunciados **[Yaneva et al. 2024]**. Uma das abordagens mais promissoras envolve o uso de *word embeddings*, que capturam propriedades semânticas das palavras por meio

de representações vetoriais. Aplicando vetores pré-treinados do tipo Word2Vec a questões de ciências da natureza, **[Jaloto et al. 2023]** obtiveram uma correlação de 0,50 entre as estimativas geradas por regressão e os valores empíricos do parâmetro b, evidenciando o potencial da técnica para aplicação educacional.

A aplicação da TRI, especialmente em sua versão de três parâmetros (3PL), é amplamente utilizada pelo INEP para calibrar os itens do ENEM **[Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira 2021]**. No entanto, esse processo depende de pré-testagens, o que eleva os custos operacionais e compromete a validade das questões, já que a exposição prévia torna sua reutilização insegura. Como alternativa, métodos baseados em Processamento de Linguagem Natural (PLN) permitem explorar a possibilidade de estimar a dificuldade de um item antes mesmo de sua aplicação, utilizando apenas seu conteúdo textual. Modelos eficientes nesse sentido possibilitariam maior controle da distribuição de dificuldade das provas, contribuindo para sua padronização e equidade.

No estudo de **[Jaloto et al. 2023]**, foram analisadas cerca de 600 questões de ciência da natureza aplicadas entre 2009 e 2020. As questões com valores extremos de dificuldade (acima de 3 ou abaixo de -3) foram descartadas, embora essa escolha não tenha sido justificada tecnicamente. Os autores utilizaram vetores de 300 dimensões, com pré-processamento textual baseado no modelo de Hartmann et al. (2017), removendo números, *stopwords* e repetições. Um modelo de regressão linear foi ajustado utilizando 80% dos dados como conjunto de treino, implementado com o pacote tidymodels. O R^2 obtido foi de 25%, embora não se tenha especificado se o valor se refere à versão ajustada da métrica.

Este trabalho amplia o escopo temporal da análise até o ano de 2023, com ênfase nas edições a partir de 2017. A partir desse ano, o ENEM passou a oferecer provas adaptadas ao formato LEDOR, que consiste na leitura oral dos enunciados para participantes com deficiência visual. Essa característica contribui para a padronização dos conteúdos imagéticos e semióticos, tornando os enunciados mais informativos e compatíveis com a vetorização textual. Dessa forma, espera-se obter representações mais consistentes para a modelagem preditiva.

Além de *embeddings* estáticos como Word2Vec **[Hartmann et al. 2017]**, o presente estudo também considera o uso de modelos contextualizados, como o BERT (Bidirectional Encoder Representations from Transformers) **[Devlin et al. 2019]**, que capturam relações semânticas levando em conta o contexto das palavras em uma sentença. Modelos desse tipo têm demonstrado avanços significativos em tarefas linguísticas diversas e oferecem uma abordagem mais robusta para representar o conteúdo dos enunciados em sua totalidade.

A modelagem preditiva é realizada por meio de regressões lineares e não lineares, visando avaliar o grau de associação entre os vetores textuais e o parâmetro de dificuldade. Modelos lineares são empregados como linha de base pela sua simplicidade e interpretabilidade, enquanto variantes mais complexas — como regressões com regularização, árvores de regressão e redes neurais — são consideradas para futuras extensões, com foco no aprimoramento do desempenho preditivo.

Também se busca compreender de que forma diferentes dimensões semânticas in-

fluenciam a dificuldade estimada. Por meio de análises de coeficientes e visualizações como nuvens de palavras, procura-se identificar padrões vocabulares associados a itens considerados fáceis ou difíceis, sejam eles de natureza técnica, científica ou cotidiana. Essa investigação pode oferecer subsídios tanto para a elaboração mais criteriosa de questões quanto para o avanço do uso de PLN em áreas ainda pouco familiarizadas com esse tipo de ferramenta, como licenciatura e pedagogia.

Em síntese, o presente estudo se propõe a reproduzir o trabalho de **[Jaloto et al. 2023]** com adaptações metodológicas, avaliando a viabilidade de prever a dificuldade de itens do ENEM exclusivamente com base em seus enunciados. A expectativa é contribuir para o desenvolvimento de métodos que tornem as provas mais homogêneas em termos de dificuldade, ao mesmo tempo em que se amplia o diálogo entre a computação e os campos da avaliação educacional.

2. Metodologia

Como base para análise, foram tomadas apenas as questões de Ciências Humanas das provas do Exame Nacional do Ensino Médio (ENEM) entre os anos de 2017 e 2023, com foco nas versões adaptadas para aplicação com recurso de leitura por ledor. A metodologia foi estruturada em uma série de *notebooks* implementados em Python. O código-fonte está publicamente disponível em `https://github.com/ciziks/enem-word-embeddings`.

2.1. Extração das Questões

O primeiro passo deste estudo corresponde à etapa mais morosa de sua execução: a extração e estruturação do conjunto de dados. Isso se deve ao fato de que, embora as provas e os microdados estejam disponibilizados pelo INEP **[Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira 2025]**, não há padronização na estrutura dos arquivos PDF. Enquanto alguns documentos estão em formato texto, outros apresentam trechos criptografados ou compostos por imagens, o que dificulta a extração automática das informações.

Com o objetivo de padronizar minimamente a extração dos enunciados, foi implementado o notebook *"1. Extração dos Enunciados.ipynb"*, que contém scripts para coleta e extração dos enunciados das questões do ENEM a partir dos cadernos oficiais. Nessa etapa, empregou-se o pacote `PyPDF2` para extrair o texto bruto (em formato de `string`) dos arquivos PDF referentes aos anos de 2017, 2018, 2019, 2020, 2022 e 2023. Ressalta-se que, no ano de 2021, não foi possível realizar a extração, pois a prova foi disponibilizada com um modelo específico de criptografia que inviabilizou a coleta automatizada. No total, foram extraídas 267 questões com suas respectivas alternativas.

Além da extração dos enunciados, foi realizado o cruzamento desses dados com os parâmetros psicométricos presentes nos microdados do INEP, com destaque para o parâmetro *b* da Teoria de Resposta ao Item (TRI), além do gabarito oficial. É importante ressaltar que, para este estudo, foram selecionadas exclusivamente as questões das provas de Ciências Humanas, em razão da riqueza e variedade de conteúdo textual, o que favorece a aplicação de métodos baseados em Processamento de Linguagem Natural (PLN).

2.2. Pré-Processamento dos Dados

No notebook *"2. Pré-Processamento.ipynb"*, foi aplicada uma sequência de transformações linguísticas aos textos dos enunciados **[Jaloto et al. 2023]**: conversão para letras minúsculas, remoção de pontuação, exclusão de *stopwords* (palavras muito frequentes que não carregam significado relevante para a análise, como "de", "e", "por"), remoção de números e lematização (redução das palavras à sua forma canônica, como transformar "correndo" em "correr").

Além disso, eliminaram-se palavras duplicadas dentro de cada item, com o intuito de reduzir o ruído nas representações vetoriais subsequentes. O objetivo dessa etapa foi padronizar os dados para a fase de vetorização, garantindo consistência na semântica dos termos representados e distinguindo adequadamente o papel de cada alternativa na estrutura da prova.

2.3. Análise Exploratória

Com o objetivo de promover uma melhor familiarização com os dados, o notebook *"3. Análise Exploratória.ipynb"* reuniu estatísticas descritivas sobre a base textual e os parâmetros de dificuldade. Foram analisadas variáveis como o tamanho dos enunciados (em número de caracteres e palavras), a distribuição do parâmetro *b* ao longo dos anos e a relação entre o comprimento dos textos e o nível de dificuldade associado.

Para essa investigação, foram utilizados histogramas, *boxplots* e gráficos de dispersão, a fim de identificar tendências, outliers e padrões preliminares presentes nos dados (Apêndice A).

2.4. Vetorização

A vetorização dos textos foi realizada no notebook *"4. Tokenização.ipynb"*. Utilizou-se, para isso, a média dos vetores de *word embeddings* pré-treinados do modelo desenvolvido pelo NILC **[Hartmann et al. 2017]** com 300, 100 e 50 dimensões, com o objetivo de representar cada item. A escolha por diferentes tamanhos de vetores visou comparar o desempenho dos modelos em função da dimensionalidade da representação vetorial.

Cada palavra foi substituída por seu vetor correspondente e, em seguida, foi calculada a média vetorial do enunciado. Essa abordagem, conhecida como *mean pooling*, foi adotada com o intuito de replicar as condições originais propostas no artigo de referência. O resultado dessa etapa foi uma matriz numérica de dimensão $n \times d$, em que n representa o número de itens e d corresponde ao número de dimensões dos *embeddings* utilizados.

A partir desta etapa, passou-se a utilizar arquivos no formato `.pkl` (*Pickled Python Objects*), dada a melhora de performance ao trabalhar com vetores — formato comumente utilizado pelo *scikit-learn*, que armazena dados de maneira mais otimizada para a linguagem Python.

2.5. Engenharia de Covariáveis com uso de LLMs

Inspirado por pesquisas recentes, o estudo utiliza *Large Language Models* (LLMs) como uma ferramenta para engenharia de atributos (ou *features*). A premissa é que o desempenho de um LLM ao tentar resolver uma questão pode servir como um *proxy* para a

dificuldade percebida do item **[Yaneva et al. 2024]**. Respostas de diferentes LLMs podem simular estudantes com vários níveis de proficiência, fornecendo uma fonte adicional de informação para a predição da dificuldade. **[Park et al. 2024]**

A partir disso, a extração dessas covariáveis foi conduzida por meio dos notebooks *"5.1 LLM Features - Extração.ipynb"* e *"5.2 LLM Features - Tratamento.ipynb"*. Essa etapa teve como objetivo complementar as análises baseadas na vetorização dos enunciados, a partir da avaliação da capacidade desses modelos em resolver questões do ENEM. Com base nos resultados obtidos por dois modelos de linguagem de larga escala amplamente acessíveis para fins acadêmicos — o **LLaMA-3.2 [Grattafiori et al. 2024]**, com taxa de acerto de 57%; e o **DeepSeek-v3 [Guo et al. 2025]**, com 82% —, foi construída uma variável binária indicativa de acerto para cada item. Essa variável visa incorporar informações semânticas adicionais à modelagem preditiva, enriquecendo as representações vetoriais e potencialmente contribuindo para a melhoria do desempenho dos modelos.

2.6. Cálculo de Similaridades entre Embeddings

Como última variável considerada, foram calculadas as similaridades de cosseno entre o enunciado da questão, o gabarito e os distratores, a partir de suas representações vetoriais (*embeddings*) com 300, 100 e 50 dimensões. O objetivo dessa abordagem foi quantificar o grau de proximidade semântica entre os diferentes elementos textuais, de modo a capturar indícios potenciais de *"atração"* ou *"confusão"* entre o enunciado e as opções de resposta. **[AlKhuzaey et al. 2024]**

Os valores de similaridade resultantes variam teoricamente no intervalo $[-1, 1]$, mas, devido à natureza dos embeddings do modelo Word2Vec utilizado — cujos vetores apresentam predominantemente componentes positivos —, os resultados observados concentram-se no intervalo $[0, 1]$. Valores mais próximos de 1 indicam maior proximidade semântica entre os textos comparados, ao passo que valores próximos de 0 refletem baixa similaridade.

2.7. Modelagem Preditiva

Nessa etapa, foram utilizados modelos de predição baseados tanto em algoritmos de aprendizado de máquina quanto em regressão linear simples, com o objetivo de estimar o parâmetro de dificuldade (b) a partir das representações vetoriais dos enunciados.

Como linha de base, foram empregados modelos de regressão linear simples, incluindo regressão linear clássica, *Ridge Regression* e *Lasso Regression*, implementados com o auxílio das bibliotecas `statsmodels` e `scikit-learn`. A regressão linear clássica estima os coeficientes por meio da minimização do erro quadrático entre os valores observados e previstos. Já as variantes Ridge e Lasso introduzem regularização na função de custo: a primeira adiciona uma penalização baseada na norma ℓ_2 dos coeficientes (evitando valores extremos), enquanto a segunda utiliza a norma ℓ_1, promovendo também a seleção de variáveis ao reduzir alguns coeficientes a zero.

Além dos modelos lineares, foram exploradas também árvores de regressão (*Decision Tree Regressors*) como abordagem alternativa. Esses modelos seguem uma estrutura hierárquica de decisões, particionando o espaço de atributos com base em divisões que minimizam o erro de predição em cada subconjunto. Por sua capacidade de modelar

relações não lineares e interações entre variáveis, as árvores oferecem maior flexibilidade em relação aos modelos lineares tradicionais.

Para a avaliação dos modelos, os dados foram divididos em conjuntos de treino e teste, na proporção de 70%-30%. O desempenho foi medido por meio da correlação entre os valores previstos e observados do parâmetro *b*, além de análises gráficas dos resíduos.

Complementarmente ao modelo linear proposto no artigo original de referência, também foi implementada uma arquitetura neural baseada no modelo pré-treinado `neuralmind/bert-base-portuguese-cased`, doravante referida como *BertimbauRegressor*. A escolha por essa abordagem se justifica pelos resultados apresentados por **[Nasir et al. 2024]**, que destacam os modelos baseados em BERT como técnica promissora para predição de dificuldade textual, especialmente em contextos educacionais.

3. Experimentos

3.1. Medida de Avaliação: $RMSE$

Como descrito em **[AlKhuzaey et al. 2024]**, a métrica de avaliação mais utilizada em tarefas de predição de dificuldade é a *Raiz do Erro Quadrático Médio* (RMSE). Ela mede o erro de previsão na mesma unidade da variável original, facilitando a interpretação da magnitude do erro absoluto [Willmott and Matsuura 2005]. Um valor menor indica maior precisão. Contudo, é uma métrica muito sensível a *outliers* devido à quadratura dos erros **[Willmott and Matsuura 2005]**. Matematicante é definida por:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\,,$$

onde y_i são os valores observados e $\hat{y}_i$ os valores preditos pelo modelo.

3.2. Modelos de Regressão com *Word Embeddings*

Nesta etapa, investigou-se a possibilidade de predizer o parâmetro de dificuldade (b) dos itens de Ciências Humanas como um problema de Regressão. Para isso, foram utilizados *embeddings* pré-treinados do tipo `Word2Vec`, aplicados ao enunciado dos itens, com três diferentes dimensionalidades: 300, 100 e 50. Para cada uma dessas representações, ajustaram-se dois modelos de regressão: regressão linear múltipla (OLS) e regressão com regularização Lasso.

Além disso, foram feitas 3 abordagens: a aplicação apenas da representação vetorial do enunciado; a exclusão de *outliers* (como sugerido em **[Jaloto et al. 2023]**); e utilizando as covariáveis extraídas dos LLMs e similaridades **[Nasir et al. 2024]**. Vale destacar também que variável resposta passou por transformação de Box-Cox, sendo necessário o deslocamento dos valores a fim de garantir sua positividade. Curiosamente, a transformação resultou no mesmo parâmetro ótimo em todos os casos ($\lambda \approx 0{,}630$), o que indica certa estabilidade na distribuição dos valores de b após o ajuste.

A Tabela 1 apresenta os resultados obtidos para cada uma das três dimensionalidades de embeddings, em todos os métodos e abordagens:

Tabela 1. Resultados RMSE dos Modelos de Regressão

Modelos		**RMSE**		
		300 dimensões	**100 dimensões**	**50 dimensões**
Regressão Linear	Enunciado Embedding	0.7721	0.7526	0.5395
	Exclusão de Outliers	1.2189	1.1116	0.7040
	Features LLM + Similaridades	0.7479	0.6304	0.5409
Regressão Lasso	Enunciado Embedding	0.5177	0.4936	0.4936
	Exclusão de Outliers	0.6342	0.6879	0.6086
	Features LLM + Similaridades	**0.4881**	**0.4936**	**0.4936**

3.3. Explorando Modelos Não-Lineares

Em uma abordagem complementar, explorou-se o uso de modelos de aprendizado de máquina mais robustos para a predição do parâmetro de dificuldade (b). Focando na representação vetorial de 50 dimensões, que apresentou um bom balanço entre complexidade e desempenho na análise anterior, foram avaliados quatro algoritmos distintos: Regressão Ridge, Árvore de Regressão, uma versão otimizada da Árvore de Regressão com busca de hiperparâmetros e, por fim, um *ensemble* de árvores com o Random Forest Regressor.

O objetivo foi verificar se modelos com maior capacidade de capturar relações não-lineares poderiam superar as abordagens lineares. O desempenho também foi medido pelo RMSE, comparando os resultados nos conjuntos de treino e teste para avaliar a capacidade de generalização e o risco de *overfitting* (overfitting) de cada modelo.

A Tabela 2 resume os valores de RMSE obtidos. Nota-se que, enquanto a Árvore de Regressão tradicional alcança o menor erro de treino, ela sofre com o maior grau de *overfitting*. A versão otimizada, por sua vez, foi capaz de mitigar esse efeito, apresentando o resultado mais promissor no conjunto de teste e demonstrando a importância do ajuste de hiperparâmetros para este tipo de modelo.

Tabela 2. Resultados dos Modelos Árvore.

	Embeddings de 50 dimensões			
	Ridge Regression	**Árvore de Regressão**	**Árvore de Regressão (Otimizada)**	**Random Forest Regressor**
RMSE Treino	0.4251	**0.2367**	0.4746	0.4579
RMSE Teste	0.4884	0.5452	**0.4550**	0.4974

3.4. BERT Regressor

Para esta tarefa, utilizamos o modelo BERT (*Bidirectional Encoder Representations from Transformers*) **[Devlin et al. 2019]**, que se destaca por sua capacidade de analisar o texto em ambas as direções, capturando um contexto mais rico que modelos anteriores. Sua arquitetura, baseada em *Transformers*, avalia a relação entre todas as palavras simultaneamente, se tornando a melhor alternativa nas últimas pesquisas na tarefa de predição de dificuldade **[Yaneva et al. 2024]**. Assim como sugerido no artigo original, para o *fine-tuning* não foram removidas stopwords, pontuações ou números do texto, mantendo o contexto original para o modelo. O BERT foi pré-treinado com tarefas de previsão de palavras mascaradas e, para este trabalho, ajustado para regressão com a adição de uma camada de saída específica.

Após o treinamento por 10 épocas, a avaliação final do modelo revelou um forte *overfitting*, não superando boa parte dos métodos de regressão aplicados anteriormente. Para investigar se um contexto de entrada mais rico poderia melhorar o desempenho, foram treinadas duas versões: a primeira, usando apenas o enunciado da questão, e a segunda, usando o enunciado junto com as alternativas. Os resultados foram muito similares e demonstraram a mesma tendência como é possível observar na Tabela 3.

Tabela 3. Resultados de RMSE para os dois modelos treinados.

Modelo	Entrada	RMSE de Treino	RMSE de Teste
BERT Regressor	Apenas o Enunciado	0.19	0.55
BERT Regressor	Enunciado + Alternativas	0.18	0.57

A performance quase idêntica, com uma leve piora no conjunto de teste ao adicionar as alternativas, evidencia que o problema central não é a falta de contexto na entrada, mas sim a escassez de dados para treinamento. Como referência, o artigo original **[Devlin et al. 2019]** sugere ao menos 100 mil exemplos para o *fine-tuning* de tarefas de regressão ou classificação. Desse modo, como trabalho futuro, é essencial ampliar a base de dados para obter resultados mais robustos. A implementação de *early-stopping* para recuperar o melhor modelo dentre as épocas, sugerido pelo professor, embora uma boa prática, não foi suficiente para contornar essa limitação fundamental. Os resultados demonstram que o modelo memorizou os poucos dados de treino, mas não aprendeu a generalizá-los. Portanto, a prioridade para trabalhos futuros é a expansão do volume de dados, permitindo que o modelo aprenda padrões robustos e alcance resultados melhores.

4. Conclusão

Os resultados obtidos ao longo deste estudo evidenciam o potencial das representações vetoriais para a predição da dificuldade de itens do ENEM na área de Ciências Humanas. Modelos lineares, como a regressão Lasso, alcançaram desempenho consistente, sobretudo quando combinados com *embeddings* de menor dimensionalidade (50 dimensões) e variáveis resultantes da similaridade semântica e respotas de LLMs. Essa abordagem foi capaz de capturar nuances relevantes do texto, permitindo a redução do erro de predição (RMSE) em comparação com modelos lineares simples baseados apenas nos enunciados. A introdução de covariáveis binárias derivadas do desempenho de LLMs, ainda, como o DeepSeek-v3 e o LLaMA-3.2, mostrou-se particularmente promissora, sugerindo que modelos de linguagem podem simular, ainda que de forma limitada, a percepção humana de dificuldade. Por outro lado, embora modelos não lineares como Árvores de Decisão otimizadas e *Random Forests* tenham apresentado avanços pontuais, seus ganhos foram limitados frente ao risco de *overfitting*. De forma semelhante, a implementação do BERT Regressor, embora metodologicamente robusta, foi prejudicada pela quantidade reduzida de dados, o que restringiu seu poder de generalização e reafirma a importância da escala em treinamentos com *transformers*.

Diante desse panorama, algumas direções futuras se impõem. Em primeiro lugar, a ampliação da base de dados é uma necessidade latente, especialmente para viabilizar o uso pleno de modelos baseados em *deep learning*, como o BERT, cujo desempenho está diretamente associado ao volume e à diversidade de exemplos durante o *fine-tuning*. Como próximos passos, propõe-se: (i) a ampliação e diversificação da base de dados, utilizando técnicas como *data augmentation* com reformulações semânticas; (ii) a exploração de *embeddings* contextuais multimodais que considerem não apenas o texto, mas também imagens ou gráficos eventualmente presentes nos itens; e (iii) a incorporação de dados empíricos complementares, como tempos de resposta médios e padrões de acerto em diferentes níveis de proficiência. Com esses avanços, espera-se consolidar uma metodologia que contribua de forma prática para a elaboração de avaliações mais balanceadas, seguras e economicamente viáveis.

Referências

AlKhuzaey, S., Grasso, F., Payne, T., and Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34:862–914.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. Acesso em: 02 jun. 2025.

Grattafiori, A., Dubey, A., and Jauhri, A. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. Acesso em: 24 mai. 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., and et al., H. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. Acesso em: 24 mai. 2025.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. http://arxiv.org/abs/1708.06025. arXiv:1708.06025 [cs], Acesso em: 30 abr. 2025.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2021). Entenda a sua nota no enem: guia do participante. https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/entenda_a_sua_nota_no_enem_guia_do_participante.pdf. Acesso em: 30 abr. 2025.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2025). Microdados do enem. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 30 abr. 2025.

Jaloto, A., Peres, A. J. d. S., Zuanazzi, A. C., Cainã, A., and Primi, R. (2023). É possível calibrar os itens do enem sem pré-teste? https://www.even3.com.br/anais/xiiabave/661223-e-possivel-calibrar-os-itens-do-enem-sem-pre-teste/. Acesso em: 30 abr. 2025.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons, Hoboken, NJ.

Nasir, M., Shaheen, E., and Hoenkamp, E. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*. Acesso em: 24 mai. 2025.

Park, J.-W., Park, S.-J., Won, H.-S., and Kim, K.-M. (2024). Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177. Association for Computational Linguistics. Acesso em: 24 mai. 2025.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean squared error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82.

Yaneva, V., North, K., Baldwin, P., Ha, L. A., Rezayi, S., Zhou, Y., Choudhury, S. R., Harik, P., and Clauser, B. (2024). Automated prediction of difficulty and response time for multi-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 1–17. Association for Computational Linguistics. Acesso em: 24 mai. 2025.

APÊNDICES

A Análise Exploratória

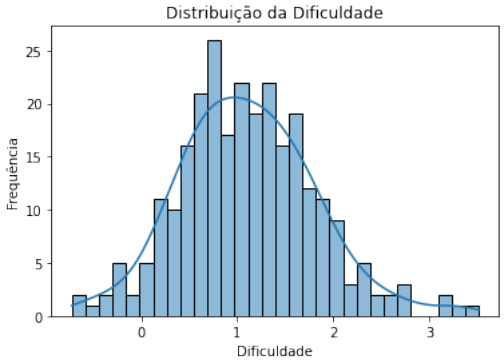

Figura 1. Distribuição do parâmetro B

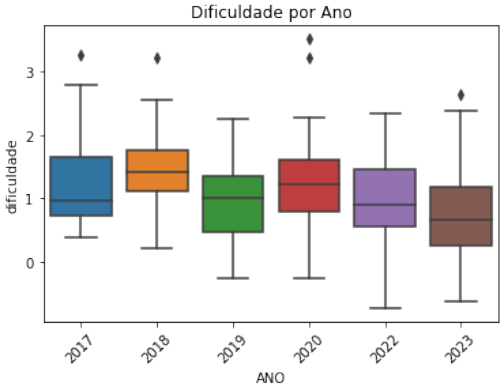

Figura 2. Distribuição do parâmetro B por ano

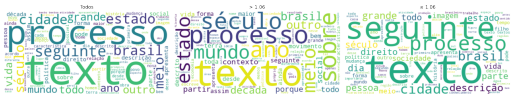

Figura 3. Nuvem de palavras do enunciado (por dificuldade)

Figura 4. Nuvem de palavras do gabrito (por dificuldade)

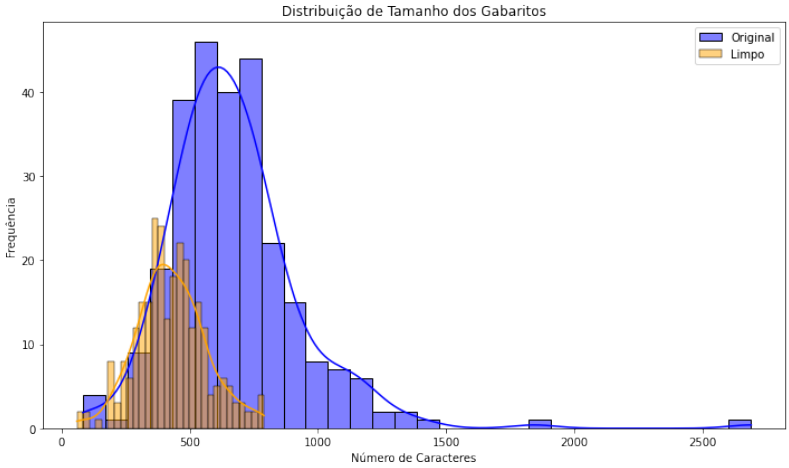

Figura 5. Distribuição do tamanho do enunciado (pré e pós tratamento)