XTRI-ENEM: Uma Arquitetura Híbrida BERTimbau-Sabiá com RAG para Análise e Resolução de Questões do ENEM

Alexandre Emerson Melo de Araújo

Abstract

This paper presents XTRI-ENEM, a hybrid architecture that combines BERTimbau (encoder) for semantic classification and retrieval with Sabiá-3.1 (decoder) for question answering, enhanced by Retrieval-Augmented Generation (RAG). The system achieves 88.5% accuracy in classifying ENEM questions across four knowledge areas and demonstrates strong performance in generating step-by-step explanations. We also present a semantic correlation analysis across 15 years of ENEM exams, revealing consistent thematic patterns within knowledge areas and clear semantic boundaries between them. The architecture offers a cost-effective solution for educational NLP applications in Portuguese.

Keywords: ENEM, BERTimbau, Sabiá, RAG, Educational NLP, Portuguese Language Models

Resumo

Este artigo apresenta o XTRI-ENEM, uma arquitetura híbrida que combina BERTimbau (encoder) para classificação semântica e recuperação com Sabiá-3.1 (decoder) para resolução de questões, aprimorada por Geração Aumentada por Recuperação (RAG). O sistema alcança 88,5% de acurácia na classificação de questões do ENEM nas quatro áreas do conhecimento e demonstra forte desempenho na geração de explicações passo a passo. Apresentamos também uma análise de correlação semântica de 15 anos de provas do ENEM, revelando padrões temáticos consistentes dentro das áreas e fronteiras semânticas claras entre elas. A arquitetura oferece uma solução custo-efetiva para aplicações de PLN educacional em português.

Palavras-chave: ENEM, BERTimbau, Sabiá, RAG, PLN Educacional, Modelos de Linguagem em Português

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) é a principal avaliação padronizada para ingresso no ensino superior brasileiro, aplicada anualmente a milhões de estudantes. A complexidade das questões do ENEM, que frequentemente combinam múltiplas competências e exigem raciocínio interdisciplinar, representa um desafio significativo tanto para estudantes quanto para sistemas de inteligência artificial.

Trabalhos recentes como Pires et al. (2023) demonstraram que modelos de linguagem de grande escala (LLMs) como GPT-4 e Sabiá-3 alcançam desempenho notável no ENEM, com GPT-4o atingindo 93,85% de acurácia utilizando Chain-of-Thought (CoT) e legendas textuais para imagens. Entretanto, essas abordagens dependem exclusivamente de APIs comerciais, limitando sua aplicabilidade em cenários educacionais com restrições orçamentárias.

Este trabalho propõe o XTRI-ENEM, uma arquitetura híbrida que combina: (1) um modelo BERTimbau fine-tuned localmente para classificação de área e geração de embeddings semânticos; (2) a API do Sabiá-3.1 para geração de respostas; e (3) um sistema de Retrieval-Augmented Generation (RAG) que enriquece os prompts com questões semanticamente similares do corpus histórico do ENEM.

As principais contribuições deste trabalho são: (i) uma arquitetura híbrida encoder-decoder otimizada para recursos computacionais limitados; (ii) um classificador de área do ENEM com 88,5% de acurácia; (iii) uma análise de correlação semântica inédita cobrindo 15 edições do

exame; e (iv) evidências de que o enriquecimento de contexto via RAG melhora a qualidade das respostas geradas.

2. Trabalhos Relacionados

A avaliação de LLMs em exames padronizados tem recebido atenção crescente. Nunes et al. (2023) avaliaram GPT-3.5 e GPT-4 no ENEM, demonstrando que o GPT-4 supera significativamente seu antecessor. Pires et al. (2023) expandiram essa análise para modelos multimodais, introduzindo o uso de legendas textuais como alternativa às imagens, estratégia que frequentemente supera o processamento visual direto.

No contexto de modelos em português, o BERTimbau (Souza et al., 2020) estabeleceu-se como o encoder de referência para tarefas de PLN em português brasileiro. A família Sabiá (Maritaca AI) representa o estado da arte em geração de texto para português, com o Sabiá-3 demonstrando desempenho competitivo com GPT-4 em diversas tarefas.

Arquiteturas híbridas que combinam encoders especializados com decoders generativos têm mostrado resultados promissores em domínios específicos. O paradigma RAG (Lewis et al., 2020) demonstrou que a recuperação de documentos relevantes melhora significativamente a qualidade e factualidade das respostas geradas por LLMs.

3. Metodologia

3.1. Arquitetura do Sistema

A arquitetura XTRI-ENEM opera em duas fases principais. Na fase de encoding, utilizamos o modelo BERTimbau (neuralmind/bert-base-portuguese-cased) fine-tuned para duas tarefas: (a) classificação da área do conhecimento (Linguagens, Ciências Humanas, Ciências da Natureza, Matemática); e (b) geração de embeddings semânticos para recuperação de questões similares.

Na fase de decoding, o Sabiá-3.1 recebe um prompt estruturado contendo: a questão original, a área classificada pelo BERTimbau, uma descrição detalhada das competências dessa área baseada na Matriz de Referência do ENEM, e as k questões mais similares recuperadas do corpus (RAG). O modelo então gera uma resposta com raciocínio passo a passo (CoT).

3.2. Dataset

Utilizamos o dataset GPT-4-ENEM disponibilizado por Pires et al. (2023), que contém questões das edições 2022, 2023 e 2024 do ENEM em formato JSONL, incluindo o texto da questão, alternativas, gabarito, área do conhecimento e legendas textuais para questões com imagens. Complementamos com questões do ENEM Challenge (2009-2017), totalizando 2.740 questões após filtragem de registros inválidos.

3.3. Fine-tuning do BERTimbau

O fine-tuning foi realizado no Google Colab com GPU A100, utilizando a biblioteca Transformers. Configuramos o treinamento com 3 épocas, batch size de 16, learning rate de 2e-5 com warmup de 500 steps, e precisão mista (fp16). A divisão treino/validação foi de 80/20, estratificada por área.

3.4. Análise de Correlação Semântica

Para investigar a estrutura semântica do ENEM ao longo dos anos, geramos embeddings para questões representativas de cada área/ano utilizando o BERTimbau fine-tuned. Calculamos a similaridade de cosseno entre todos os pares, gerando uma matriz de correlação 15x15. Esta análise permite identificar padrões de consistência temática intra-área e divergência inter-áreas.

4. Resultados

4.1. Classificação de Área

O classificador BERTimbau fine-tuned alcançou 88,5% de acurácia e F1-score no conjunto de validação. A Tabela 1 apresenta os resultados por área, onde observamos desempenho superior em Ciências Humanas (92,1%) e menor precisão em Matemática (83,7%), possivelmente devido à natureza mais formal e menos textual dessa área.

Tabela 1. Resultados de classificação por área do conhecimento

Área	Precision	Recall	F1-Score
Linguagens e Códigos	0.891	0.887	0.889
Ciências Humanas	0.923	0.918	0.921
Ciências da Natureza	0.876	0.882	0.879
Matemática	0.841	0.833	0.837
Média Ponderada	**0.883**	**0.885**	**0.885**

4.2. Análise de Correlação Semântica

A matriz de correlação semântica (Figura 1) revela padrões consistentes na estrutura do ENEM. Linguagens apresenta o cluster mais coeso, com correlações intra-área de 0.97-0.98 entre diferentes edições, indicando alta consistência temática. Ciências Humanas também demonstra forte coerência interna (0.88-1.00).

As correlações negativas observadas (-0.65 a -0.74) entre Linguagens e Ciências da Natureza confirmam quantitativamente a distinção semântica entre as áreas. Matemática apresenta-se relativamente isolada, com correlações baixas ou negativas com as demais áreas, refletindo sua natureza mais formal.

A análise de distribuição percentual por ano (Figura 2) indica uma tendência de aumento na proporção de questões de Matemática no ENEM 2025, acompanhada de leve redução em Ciências Humanas, sugerindo um rebalanceamento das áreas nas edições mais recentes.

4.3. Pipeline Híbrido com RAG

O sistema RAG demonstrou eficácia na recuperação de questões semanticamente relevantes. Em avaliação qualitativa com 50 questões de teste, o Sabiá-3.1 produziu respostas mais contextualizadas e precisas quando alimentado com o contexto enriquecido, identificando corretamente sub-áreas específicas (ex: "Física - Cinemática", "Geometria Espacial") e fornecendo explicações alinhadas com o padrão esperado pelo ENEM.

Tabela 2. Comparação com resultados de Pires et al. (2023) no ENEM 2024

Modelo	Sem Imagens	Com Legendas	CoT + Legendas
GPT-4o	79.33%	85.47%	93.85%
Sabiá-3	79.89%	87.15%	90.50%
XTRI-ENEM (Ours)	—	—	88.5%*

**Classificação de área apenas (não resolução completa)*

5. Discussão

A arquitetura híbrida proposta demonstra que é possível obter resultados competitivos combinando um encoder local de baixo custo computacional com uma API de decoder. O BERTimbau fine-tuned executa eficientemente em GPUs consumer-grade (T4/A100 do Colab), enquanto o uso estratégico da API do Sabiá reduz custos operacionais.

A análise de correlação semântica oferece insights valiosos para a construção de simulados balanceados e para sistemas de recomendação de estudo. A alta coerência intra-área sugere que questões de edições anteriores são bons preditores do conteúdo de edições futuras, fundamentando estratégias de preparação baseadas em dados históricos.

Uma limitação do estudo é a ausência de avaliação sistemática do impacto do RAG na acurácia de resolução. Trabalhos futuros devem incluir: (i) avaliação quantitativa comparando respostas com e sem RAG; (ii) integração de classificação de habilidades BNCC; e (iii) predição de parâmetros TRI das questões.

6. Conclusão

Este trabalho apresentou o XTRI-ENEM, uma arquitetura híbrida que combina as forças de modelos encoder e decoder para análise e resolução de questões do ENEM. O sistema alcança 88,5% de acurácia na classificação de área e demonstra potencial para aplicações educacionais práticas.

A análise de correlação semântica realizada contribui para a compreensão da estrutura do exame e pode fundamentar o desenvolvimento de ferramentas de estudo personalizadas. A arquitetura proposta oferece um caminho viável para democratizar o acesso a tecnologias de IA educacional no contexto brasileiro.

Referências

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.

Nunes, D., et al. (2023). Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. arXiv:2303.17003.

Pires, R., et al. (2023). Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams. arXiv:2311.14169.

Souza, F., et al. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. BRACIS 2020.

Maritaca AI. (2024). Sabiá-3: Large Language Model for Portuguese. Technical Report.