AVALIAÇÃO DO DESEMPENHO DE MODELOS LLAMA E GEMINI NA CORREÇÃO DE REDAÇÕES DO ENEM

Introdução

O trabalho avalia o desempenho dos modelos de linguagem LLaMA e Gemini na correção automática de redações do ENEM (Exame Nacional do Ensino Médio), com foco na atribuição de notas para as cinco competências exigidas. O estudo busca reduzir subjetividade e aumentar a eficiência no processo avaliativo, propondo inovações como engenharia de prompts e análise com aprendizado supervisionado.

Ferramentas

Modelos:

- **LLaMA 3.1 70B (Large Language Model Meta AI):** Modelo de linguagem com 70 bilhões de parâmetros e otimizado para o processamento de longas sequências de tokens.
- **Gemini 1.5-Flash:** Modelo multimodal da Google, focado em eficiência computacional e análise de dados complexos.
- **NILC-Metrix:** Ferramenta para analisar características textuais, como coesão e complexidade, oferece métricas para avaliação automatizada.
- **XGBoost (Extreme Gradient Boosting):** Algoritmo baseado em árvores de decisão, eficiente para tarefas de classificação.

Métricas:

- **QWK (Quadratic Weighted Kappa)**: Mede a concordância ponderada em dados ordinais. (1 - Concordância Perfeita | 0 - Acaso)
- **RMSE (Root Mean Square Error):** Avalia o erro médio quadrático entre valores previstos e observados.

Datasets:

- **Dataset Estendid**o: Combina o **AES ENEM Dataset** e o novo **Dataset de Redações Nota Mil,** criado neste trabalho com textos de pontuação máxima no ENEM coletados de fontes públicas. Esse corpus balanceia a escassez de redações com notas altas, sendo disponibilizado como benchmark para futuras pesquisas em avaliação automática de textos.

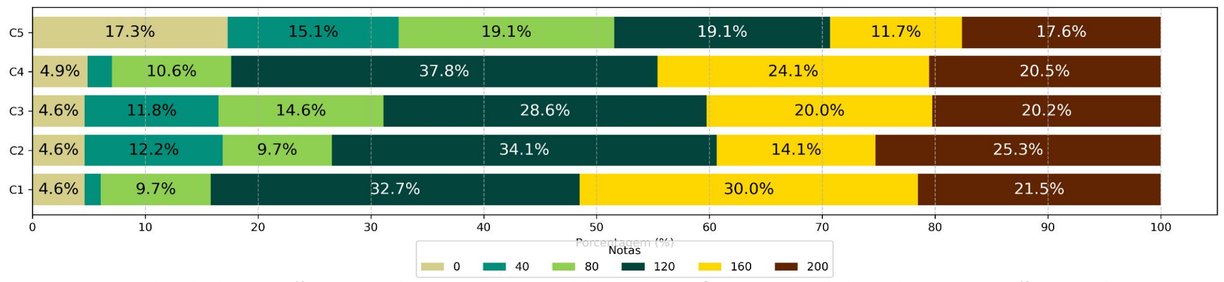

Figura 1: Distribuição de notas por competência: AES ENEM Dataset + Redações Nota Mil

Metodologia

Fase 1: Avaliação de Padrões de Prompt

Seis padrões de prompts foram avaliados para entender o impacto na precisão das notas atribuídas pelos modelos:

1. **Prompt Base:** Formato simples, pede notas numéricas sem contexto.
2. **Prompt Base em Inglês:** Tradução do prompt base, investigando se o idioma melhora desempenho dos modelos multilíngues.
3. **Prompt Melhorado:** Estruturado com ajuda do PromptPerfect, otimizando clareza e formato esperado.
4. **Prompt Persona:** Configura o modelo para atuar como um professor especializado na correção do ENEM.
5. **Prompt Cadeia de Pensamento:** Divide a tarefa em etapas, como explicação do método de avaliação, correção detalhada e geração de nota final.
6. **Prompt Contextualizado:** Adiciona descrições detalhadas das competências do ENEM e seus critérios.

Fase 2: Avaliação de Extração de Características

Os melhores resultados da Fase 1 foram usados como insumo para o modelo XGBoost, combinando as saídas dos modelos LLaMA e Gemini com 72 métricas extraídas pelo NILC-Metrix. Este modelo supervisionado foi configurado para capturar relações complexas entre características textuais e notas, resultando em previsões mais robustas.

Resultados

Fase 1: Os prompts mais detalhados **(Cadeia de Pensamento e Contextualizado) apresentaram os melhores desempenhos**, aumentando a precisão e consistência das notas atribuídas pelos modelos. O modelo **Gemini 1.5-Flash demonstrou superioridade geral**, destacando-se pela consistência e robustez em todas as competências e métricas avaliadas.

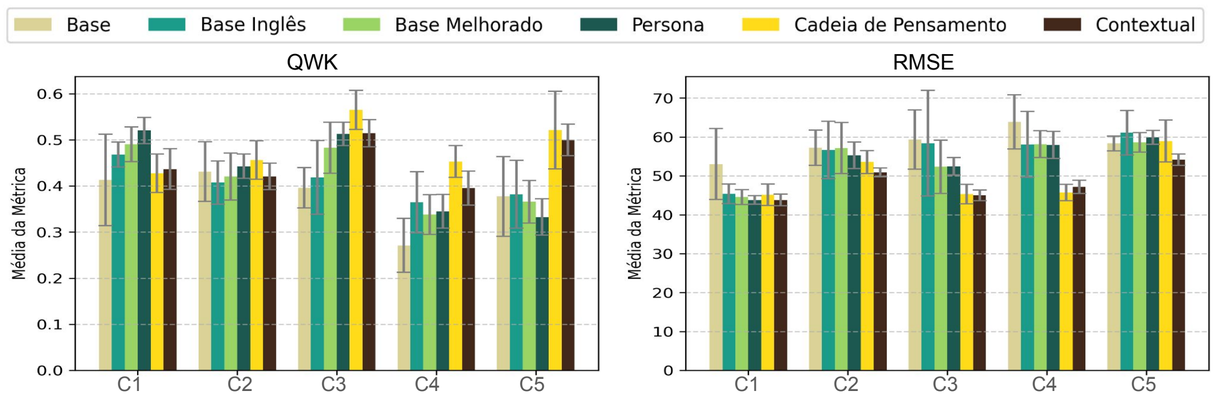

Figura 2: Resultados Gemini - QWK e RMSE - Média e Desvio 20 execuções por redação.

O LLaMA teve desempenho variável, com bons resultados no experimento Contextual em competências 1, 2 e 3. Ambos os modelos apresentaram **fraco desempenho com o experimento Base Inglês**, evidenciando a importância de prompts alinhados ao contexto do ENEM.

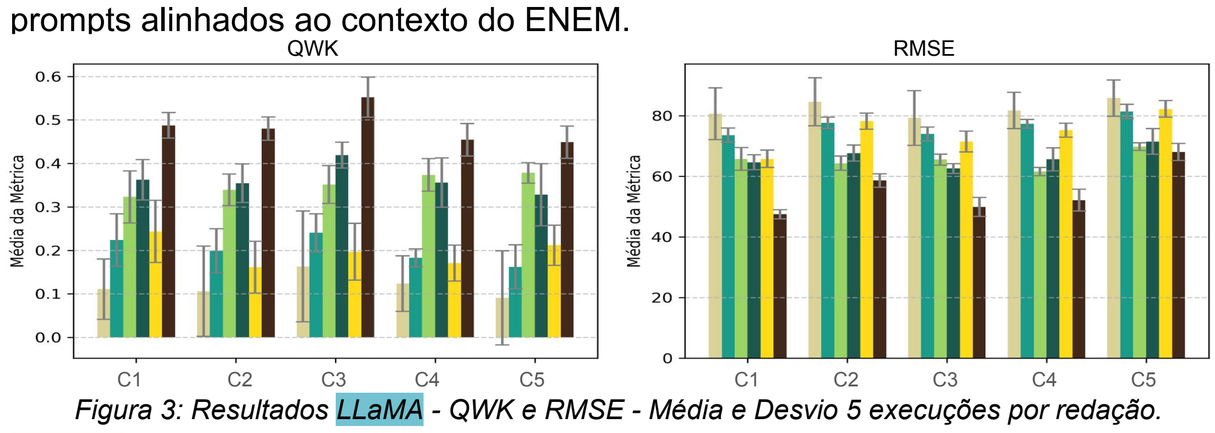

Figura 3: Resultados LLaMA - QWK e RMSE - Média e Desvio 5 execuções por redação.

Fase 2: O modelo XGBoost, treinado com métricas do NILC-Metrix e saídas dos modelos LLaMA e Gemini, **apresentou desempenho superior em quase todas as competências, destacando-se no Experimento 4 (NILC + Gemini + LLaMA)** com altos valores de QWK, em C2 e C3. Mesmo sem as métricas tradicionais, o caso 7 (Gemini + LLaMA) alcançou resultados robustos, como QWK de 0.619 em C5, indicando o valor das características dos LLMs.

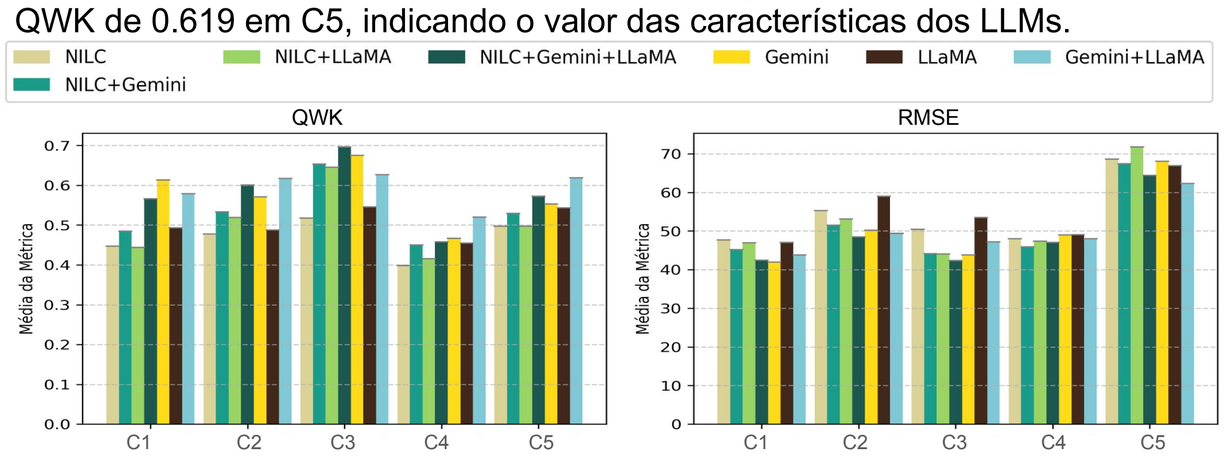

Figura 4: Resultados XGBoost com diferentes dados de treinamento com NILC, LLaMA e Gemini.

Conclusão

Este trabalho demonstrou que ajustes estratégicos em prompts e a integração de modelos de linguagem com algoritmos supervisionados melhora a avaliação automática de redações - Melhores resultados em QWK e RMSE. A inclusão de prompts estruturados, como os contextualizados e em cadeias de pensamento, mostrou-se eficaz para gerar saídas mais precisas (~↑20% QWK). A construção de um dataset com redações nota mil e a exploração de aprendizado multimodal oferecem novas oportunidades para avanços futuros na área.

Bibliografia

[al. 2023] Sidney Leal et al. ***NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese.*** *2023. url: https://arxiv.org/pdf/2201.03445.pdf*

[Dubey et al. 2024] Abhimanyu Dubey et al. ***"The llama 3 herd of models"***. *Em: arXivpreprint arXiv:2407.21783 (2024)*

[Mizumoto e Eguchi 2023] Atsushi Mizumoto e Masaki Eguchi. ***"Exploring the potential of using an ai language model for automated essay scoring"***. *Em: Research Methods in Applied Linguistics 2.2 (2023), pg. 10005*

[Silveira et al. 2024] Igor Cataneo Silveira, André Barbosa e Denis Deratani Mauá. **"A new benchmark for automatic essay scoring in Portuguese"**. Em: Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. Association for Computational Lingustics, mar. de 2024, pgs. 228–237. url: https: //aclanthology.org/2024.propor-1.23

[Team, Georgiev et al. 2024] Gemini Team, Petko Georgiev et al. ***"Gemini 1.5: unloc-king multimodal understanding across millions of tokens of context"***. *Em: arXivpreprint arXiv:2403.05530 (2024)*

[White et al. 2023] Jules White et al. ***"A prompt pattern catalog to enhance prompt engineering with chatgpt"***. *Em: arXiv preprint arXiv:2302.11382 (2023) (citado nas pgs. 3, 18).*

Autor: Daniel Silva Lopes da Costa
Orientadores: Msc. Igor Silveira | Prof. Dr. Denis Mauá
Contato: E-mail: dlscosta2016@usp.br Tel: (11) 980636381
Trabalho Completo: https://linux.ime.usp.br/~costa/tcc.html