Tucano em Question Answering: Exploração das capacidades do modelo no dataset FairytaleQA

Allan D. Ehlert[1], **Larissa A. de Freitas**[1], **Ulisses B. Corrêa**[1]

[1]Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPel)
96010-610 - Pelotas - RS - Brazil

{adehlert, larissa, ulisses}@inf.ufpel.br

Abstract. *Large Language Models (LLMs) have demonstrated remarkable capabilities in addressing various Natural Language Processing (NLP) tasks, including Question Answering (QA). To help bridge the gap between high-resource and low-resource languages, Tucano, a series of decoder-transformer models trained in Portuguese was developed. This study evaluates the performance of these models on the FairytaleQA-Translated dataset and compares results with previously tested datasets. Our findings highlight the potential of smaller, more efficient models for QA in portuguese and contribute to advancing research in low-resource languages.*

Resumo. *Modelos de Lingugaem de Grande Escala (LLMs) têm demonstrado capacidades notáveis em diversas tarefas de Processamento de Lingugaem Natural (PLN), incluindo a Geração Automática de Respostas (QA). Para ajudar a reduzir a lacuna entre idiomas de alto e baixo recurso, foi desenvolvida a série Tucano, composta por modelos decoder-transformers treinados em português. Este estudo avalia o desempenho desses modelos no dataset FairytaleQA-Translated e compara os resultados com datasets previamente testados. Nossos achados destacam o potencial de modelos menores e mais eficientes para QA em português e contribuem para o avanço da pesquisa em idiomas de baixo recurso.*

1. Introdução

As LLMs têm se destacado pela capacidade de lidar com múltiplos desafios em Processamento da Linguagem Natural (PLN), incluindo a Geração Automática de Perguntas e Respostas (QA – do Inglês Question Answering) [Wei et al. 2022]. Apesar dos avanços, ainda existe uma lacuna significativa entre os idiomas com muitos recursos e aqueles com poucos, como o português. Com o objetivo de reduzir essa desigualdade, foi desenvolvida a série Tucano [Corrêa et al. 2025], composta por modelos decoder-transformers treinados nativamente em português, variando de 160 milhões a 2,4 bilhões de parâmetros.

Este trabalho avalia o desempenho desses modelos na tarefa de QA utilizando o dataset FairytaleQA-Translated [Leite et al. 2024], assim como realiza comparações com outros *datasets* e modelos anteriormente testados, a fim de contribuir para o avanço da pesquisa em idiomas de poucos recursos e mostrando o potencial de modelos menores e mais eficientes no contexto do português.

A estrutura do trabalho está organizada da seguinte maneira: a seção de **Referencial Teórico** define conceitos essenciais para a compreensão do artigo; a seção **Trabalhos**

Relacionados apresenta a literatura relevante já publicada; a seção **Metodologia** descreve os procedimentos adotados para os experimentos; a seção **Resultados e Discussão** apresenta os resultados com as métricas utilizadas. Por fim, a seção **Conclusões** sintetiza os resultados alcançados e discute possíveis direções para estudos futuros.

2. Referencial Teórico

Modelos de Linguagem de Grande Escala (LLMs - do inglês *Large Language Models*) são tipos de modelos de linguagem desenvolvidos para produzir texto de maneira eficiente em diversas aplicações, utilizando grandes quantidades de dados durante o treinamento. Estes modelos se destacam por sua capacidade de aprender representações linguísticas a partir de imensas quantidades de dados textuais, permitindo uma alta adaptabilidade para vários tipos de tarefas, o que as tornam ferramentas poderosas para várias aplicações [Chang et al. 2023].

QA é uma tarefa de PLN, cujo objetivo é responder a perguntas formuladas em linguagem natural a partir de um texto ou base de conhecimento. Com a criação do SQuAD (*Stanford Question Answering Dataset*) [Rajpurkar et al. 2016], a tarefa de QA se consolidou como um *benchmark* imprescindível para avaliação de modelos, com uso de métricas como *Exact Match* (EM) e *F1-score*.

Few-shot learning é um método de treinamento onde um modelo é ajustado para executar uma tarefa com poucos exemplos de treinamento. Essa estratégia se posiciona entre o *fine-tuning* completo e o *zero-shot*, equilibrando desempenho e custo computacional.

Grande parte da pesquisa em torno de QA é voltada para línguas de alto recurso, como o inglês. Em contrapartida, línguas de baixo recurso possuem menos suporte e dados disponíveis, gerando uma lacuna significativa na área. Embora trabalhos recentes, como [da Rocha Junqueira et al. 2024], explorem técnicas para mitigar essa escassez, ainda há desafios relacionados à diversidade linguística do português brasileiro.

3. Trabalhos Relacionados

[Rodrigues et al. 2023] apresentaram o Albertina PT-*, um modelo de linguagem desenvolvido especialmente para aprimorar o processamento neural tanto para o português europeu quanto para o português brasileiro. O modelo utiliza como base a arquitetura DeBERTa e foi pré-treinado em um conjunto de corpora em língua portuguesa.

O trabalho de [Pires et al. 2023] propôs o modelo Sabiá, baseado em dados derivados do corpus ClueWeb2022 [Overwijk et al. 2022]. Os modelos Sabiá foram treinados em um *dataset* português usando os *frameworks* T5X e SeqIO, contabilizando 10,4 bilhões de *tokens*. A avaliação foi conduzida com o *benchmark* Poeta [Pires et al. 2023], voltado a tarefas em português, com a abordagem *few-shot*. Os exemplos foram selecionados manualmente e inseridos dentro do limite de 2.048 *tokens* por contexto. Posteriormente, o estudo de [Da Rocha Junqueira et al. 2024] analisa o modelo Sabiá-7B em várias tarefas de PLN, incluindo QA. O estudo utilizou similarmente a abordagem *few-shot*, utilizando o *dataset* SQuAD v1.1-PT para a tarefa de QA. Nestes experimentos, o modelo obteve 0,54 na métrica *F1-score* e alcançou 39% na métrica EM.

Mais recentemente, [Corrêa et al. 2025] apresentaram o Tucano, que inclui a

criação de um novo corpora para o português, o GigaVerbo, além de treinamento de modelos do tipo *decoder-transformer*. Os modelos demonstraram desempenho equivalente ou superior aos modelos existentes em múltiplos *benchmarks*.

4. Metodologia

Neste trabalho, realizamos o pré-processamento das amostras *few-shot* do *dataset* FairytaleQA-Translated, selecionando e organizando exemplos dentro das limitações de contexto do modelo. Foi montado um conjunto de treinamento adequado para QA, maximizando o número de exemplos que coubessem junto à instância de teste dentro da janela de contexto, garantindo uma representação equilibrada. Os exemplos foram então utilizados como *prompts* durante a inferência nos modelos Tucano, incluindo o componente instrucional. Após a execução, foi realizada uma análise detalhada da saída do modelo, avaliando métricas de desempenho e comparando os resultados com experimentos anteriores. O objetivo final desta análise é investigar a eficácia da abordagem *few-shot* e identificar diferenças entre *datasets*, bem como limitações e pontos fortes dos modelos.

Para avaliar o desempenho dos modelos, utilizamos duas métricas: *Exact Match* (EM) e *F1-score*. Essas métricas fornecem medidas de precisão e qualidade das respostas geradas pelos modelos. A EM é uma métrica que avalia se a resposta gerada pelo modelo corresponde exatamente à resposta correta correspondente no *dataset*, enquanto a *F1-score* é uma métrica de precisão ponderada, que combina precisão e *recall* em uma única medida, possibilitando assim melhor captação de respostas parcialmente corretas.

5. Resultados e Discussão

Foram realizados experimentos com três modelos da família Tucano, avaliados em dois conjuntos de dados de QA: SQuAD v1.1-PT e FairytaleQA. No *dataset* SQuAD v1.1-PT, observa-se que o modelo Tucano-2b4-Instruct apresentou os maiores valores, com *F1-score* de 0,15 e EM de 7,03%, enquanto as variantes menores atingem valores mais baixos. De forma semelhante, no FairytaleQA-Translated, o modelo maior se destaca, alcançando um *F1-score* de 0,21 e EM de 3,60%, com a mesma tendência de valores baixos nos modelos menores.

Para fins de comparação, dados reportados na literatura de modelos amplamente utilizados no português também foram considerados. O Albertina Base obteve *F1-score* de 0,57 e EM de 45,12%, enquanto o Albertina Large, apesar de um *F1-score* mais baixo de 0,32, alcançou 47,30% em EM. Por fim, o Sabiá-7B obteve 0,54 em *F1-score* e 39,17% em EM.

Tabela 1. Resultados dos modelos Tucano na tarefa de QA

Dataset	Modelo	F1-score	Exact Match (%)
SQuAD v1.1-PT	Tucano-160m	0,05	1,07
	Tucano-630m	0,04	0,73
	Tucano-2b4-Instruct	0,15	**7,03**
FairytaleQA-Translated	Tucano-160m	0,07	0,10
	Tucano-630m	0,10	0,33
	Tucano-2b4-Instruct	**0,21**	3,60

6. Conclusões

Os experimentos demonstraram que o desempenho dos modelos Tucano na tarefa de QA ainda é limitado no contexto de *few-shot learning*. Embora o Tucano-2b4-Instruct, o maior modelo da série, apresente ganhos consistentes em relação às suas versões menores, seus resultados permanecem inferiores aos de outros modelos, como o Sabiá-7B. Além disso, a diferença notada entre *datasets* mostra que o FairytaleQA é ainda mais desafiador para os modelos Tucano, reforçando que a adaptação de modelos de linguagem para QA em português demanda outras estratégias, como *fine-tuning* e melhor processamento dos dados de treinamento. Com isso, conclui-se que a família Tucano tem potencial promissor para aplicações em português, mas ainda enfrenta alguns obstáculos em tarefas de QA, sendo de grande importância a investigação futura dos modelos com técnicas de *fine-tuning* e métodos de *prompting* mais robustos.

Referências

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2023). A survey on evaluation of large language models.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2025). Tucano: Advancing neural text generation for portuguese. *Patterns*, page 101325.

da Rocha Junqueira, J., Freitas, L., and Corrêa, U. B. (2024). Transformer models for brazilian portuguese question generation: An experimental study. *The International FLAIRS Conference Proceedings*, 37.

Da Rocha Junqueira, J., Lopes, P., Da S. M., C. L., Silva, F. L. V., Carvalho, E. A., Freitas, L., and Brisolara, U. (2024). Sabiá in action: An investigation of its abilities in aspect-based sentiment analysis, hate speech detection, irony detection, and question-answering. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Leite, B., Osório, T. F., and Cardoso, H. L. (2024). Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages.

Overwijk, A., Xiong, C., Liu, X., VandenBerg, C., and Callan, J. (2022). Clueweb22: 10 billion web documents with visual and semantic information.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). [inline-graphic not available: see fulltext] sabiá: Portuguese large language models. In *Intelligent Systems: 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25–29, 2023, Proceedings, Part III*, page 226–240, Berlin, Heidelberg. Springer-Verlag.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). *Advancing Neural Encoding of Portuguese with Transformer Albertina PT-**, page 441–453. Springer Nature Switzerland.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models.