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

ALLAN DUARTE EHLERT[1]; LARISSA ASTROGILDO DE FREITAS[2]; ULISSES BRISOLARA CORRÊA[3]

[1]*Universidade Federal de Pelotas – adehlert@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – ulisses@inf.ufpel.edu.br*

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, OSÓRIO & CARDOSO, 2024), assim como realiza uma série de comparações com outros *datasets* anteriormente testados com estes modelos, 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, ou *low-resource languages*, possuem menos suporte e dados disponíveis, gerando uma lacuna significativa na área. Embora trabalhos recentes, como DA ROCHA JUNQUEIRA, CORRÊA e FREITAS (2024), explorem técnicas para mitigar essa escassez, ainda há desafios relacionados à diversidade linguística do português.

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), e desenvolvido a partir de três arquiteturas: LLaMA-7B, LLaMA-65B (TOUVRON et al., 2023) e GPT-J (WANG; KOMATSUZAKI, 2021). 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, CÔRREA 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*. As comparações realizadas em diferentes *benchmarks* mostraram que os modelos Tucano alcançam desempenho equivalente ou superior aos modelos já existentes.

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 (160m, 630m e 2b4-Instruct), 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 *dataset* 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.

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

Tabela 1: Comparação de performance de diferentes modelos na tarefa de QA.

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 (Tucano-160m e Tucano-630m), seus resultados permanecem inferiores aos de modelos maiores, como o Sabiá-7B. Além disso, a diferença notada entre os *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.

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.

7. REFERÊNCIAS BIBLIOGRÁFICAS

WEI, J. et al. **Emergent abilities of large language models**. *arXiv preprint*, Ithaca, 2022. Disponível em: https://arxiv.org/abs/2206.07682. Acesso em: 27 ago. 2025.

CORRÊA, U. B. et al. **Tucano: advancing neural text generation for Portuguese**. *arXiv preprint*, Ithaca, 2024. Disponível em: https://arxiv.org/abs/2411.07854. Acesso em: 27 ago. 2025.

LEITE, J.; OSÓRIO, T.; CARDOSO, H. **FairytaleQA translated: enabling educational question and answer generation in less-resourced languages**. *arXiv preprint*, Ithaca, 2024. Disponível em: https://arxiv.org/abs/2406.04233. Acesso em: 27 ago. 2025.

CHANG, J. et al. **A survey on evaluation of large language models**. *arXiv preprint*, Ithaca, 2023. Disponível em: https://arxiv.org/abs/2307.03109. Acesso em: 27 ago. 2025.

RAJPURKAR, P. et al. **SQuAD: 100,000+ questions for machine comprehension of text**. *arXiv preprint*, Ithaca, 2016. Disponível em: https://arxiv.org/abs/1606.05250. Acesso em: 27 ago. 2025.

OVERWIJK, A. et al. **ClueWeb22: 10 billion web documents with visual and semantic information**. *arXiv preprint*, Ithaca, 2022. Disponível em: https://arxiv.org/abs/2211.15848. Acesso em: 27 ago. 2025.

TOUVRON, H. et al. **LLaMA 2: open foundation and fine-tuned chat models**. *arXiv preprint*, Ithaca, 2023. Disponível em: https://arxiv.org/abs/2307.09288. Acesso em: 27 ago. 2025.

DA ROCHA JUNQUEIRA, J.; CORRÊA, U. B.; FREITAS, L. **Transformer models for Brazilian Portuguese question generation: an experimental study**. In: INTERNATIONAL FLAIRS CONFERENCE, 37., 2024. Proceedings [...]. [S. l.]: FLAIRS, 2024. v.37, n.1. DOI: 10.32473/flairs.37.1.135334. Disponível em: https://journals.flvc.org/FLAIRS/article/view/135334. Acesso em: 27 ago. 2025.

RODRIGUES, J.; GOMES, L.; SILVA, J.; BRANCO, A.; SANTOS, R.; CARDOSO, H. L.; OSÓRIO, T. **Advancing neural encoding of Portuguese with transformer Albertina PT-***. Cham: Springer Nature Switzerland, 2023. p. 441–453. DOI: https://doi.org/10.1007/978-3-031-49008-8_35.

PIRES, R.; ABONIZIO, H.; ALMEIDA, T. S.; NOGUEIRA, R. **Sabiá: Portuguese large language models**. In: NALDI, M. C.; BIANCHI, R. A. C. (Eds.). Intelligent Systems. Cham: Springer Nature Switzerland, 2023. p. 226–240.

WANG, B.; KOMATSUZAKI, A. **GPT-J-6B: a 6 billion parameter autoregressive language model**. 2021. Disponível em: https://github.com/kingoflolz/mesh-transformer-jax. Acesso em: 27 ago. 2025.