Modelos Quantizados para Question Answering em Português Brasileiro: Um Estudo Experimental com Retrieval-Augmented Generation

Júlia da Rocha Junqueira[1], **Larissa A. de Freitas**[2], **Ulisses Brisolara Corrêa**[2], **Viviane Moreira**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPel)
Pelotas – RS – Brazil

{jrjunqueira,viviane}@inf.ufrgs.br,{larissa,ulisses}@inf.ufpel.edu.br

Abstract. *This study evaluates the Sabiá-7B model on Portuguese question answering, applying quantization, fine-tuning, and retrieval-augmented generation. Results based on ROUGE metrics indicate that, despite its potential, fine-tuning with this technique reduced performance, highlighting difficulties in effectively integrating retrieved contexts, particularly in quantized models.*

Resumo. *Este trabalho investiga o desempenho do modelo Sabiá-7B na tarefa de resposta à perguntas em língua portuguesa, aplicando quantização, ajuste fino e geração de recuperação aumentada. Resultados baseados nas métricas ROUGE indicam que, apesar de seu potencial, o ajuste fino com essa técnica reduziu o desempenho, destacando dificuldades na integração eficaz dos contextos recuperados, particularmente em modelos quantizados*

1. Introdução

A tarefa de *Question Answering* (QA) envolve fornecer respostas precisas e contextualmente relevantes a consultas com base em um dado texto [Allam and Haggag, 2012], exigindo que modelos de linguagem compreendam profundamente estruturas gramaticais, identifiquem informações relevantes e as transformem em respostas coerentes [Pan et al., 2019]. Neste contexto, a necessidade de estratégias eficazes para otimizar a geração automática de respostas em modelos de linguagem natural se torna evidente.

Este trabalho investiga os efeitos do uso do método de Retrieval-Augmented Generation (RAG) [Lewis et al., 2020] em modelos de linguagem de grande escala quantizados. Em particular, exploramos o comportamento do modelo Sabiá-7B quantizado para a tarefa QA utilizando RAG. Este estudo complementa o trabalho de Yazan et al. [2024], que investigou os efeitos de quantização e RAG em modelos da língua inglesa. Os resultados obtidos demonstram que a integração do RAG no modelo quantizado compromete o desempenho, reforçando a sensibilidade à ampliação de contexto.

2. Fundamentação Teórica e Trabalhos Relacionados

A crescente acessibilidade de tecnologias baseadas em modelos de linguagem de grande escala e sua integração em aplicações práticas evidenciam sua relevância no campo do

PLN. Esse impacto é especialmente notável em tarefas como QA, que beneficia-se do avanço nos métodos de ajuste fino e técnicas para lidar com as particularidades dos dados textuais.

RAG [Lewis et al., 2020] é uma técnica que utiliza modelos probabilísticos que combinam memória paramétrica e não-paramétrica. Os modelos de RAG são construidos utilizando um modelo Transformer *seq2seq* (por exemplo, BART, FLAN-T5, T5) como memória paramétrica, e um índice denso de vetores da *Wikipedia* como memória não-paramétrica, acessado por um *retriever* neural pré-treinado.

O Sabiá-7B [Pires et al., 2023] é um modelo de linguagem de grande escala especializado para o idioma português. Este modelo foi criado para aprimorar a compreensão das estruturas linguísticas, nuances culturais e conhecimento específico do português brasileiro. Ao focar exclusivamente no português do Brasil, o modelo consegue abordar detalhes e peculiaridades que poderiam ser perdidos em um treinamento que envolvesse múltiplos idiomas.

Yazan et al. [2024] analisam o impacto da quantização em modelos de linguagem de pequeno porte no contexto de RAG. O estudo foca em modelos de 7 e 8 bilhões de parâmetros, investigando como a quantização pós-treinamento afeta suas capacidades de raciocínio em contexto longo. Os resultados indicam que a quantização pode ser aplicada com eficiência em LLMs menores, sem comprometer significativamente sua capacidade em tarefas específicas. Entretanto, em modelos como o LLaMA-2, o impacto da quantização é mais evidente, especialmente quando o tamanho do contexto aumenta devido ao maior número de documentos recuperados.

3. Metodologia

Dadas as definições apresentadas na seção anterior, descreveremos a seguir o procedimento utilizado neste estudo. Inicialmente, o conjunto de dados Pirá [Pirozelli et al., 2024] foi selecionado como base dos experimentos. Dentre os diferentes *benchmarks* do *dataset*, foi utilizada a versão padrão, já que o formato de perguntas esperado eram perguntas abertas. O seu tamanho total é de 2.2 mil instâncias, sendo 1.8 mil de treino, 225 de validação, e 227 de teste.

A partir disso, quantizamos o modelo Sabiá-7B em 4 *bits*, utilizando representação *float* de 4 dígitos (nf4) e tipo de dado *float* de 16 *bits*. Esse equilíbrio mantém precisão enquanto reduz recursos computacionais. Em conjunto com a quantização, utilizamos o método QLoRA [Dettmers et al., 2024] para reduzir o uso de memória durante o FT. Os parâmetros treináveis (3.89% do total) foram configurados com *rank*=1, *alpha*=4 e *dropout*=0.1.

Como fonte de memória não-paramétrica, utilizamos o dataset da Wikipedia em português[1], processado com *embeddings* do modelo BERTimbau[2] e disponibilizado no Hugging Face[3]. Os vetores são indexados com FAISS [Johnson et al., 2019] para buscas eficientes. A inferência combina uma frase auxiliar, o contexto do dataset, o contexto recuperado pelo RAG e a pergunta.

[1]https://huggingface.co/datasets/TucanoBR/wikipedia-PT

[2]https://huggingface.co/ricardo-filho/bert-base-portuguese-cased-nli-assin-2

[3]https://huggingface.co/datasets/jjuliar/wikipedia-PT

Por fim, as respostas geradas são submetidas a uma avaliação utilizando métricas quantitativas de ROUGE [Lin, 2004] que medem a qualidade, similaridade e precisão das respostas. O ROUGE é uma métrica clássica e amplamente utilizada em tarefas de geração de texto, especialmente em QA. Usá-lo facilita a comparação direta com trabalhos anteriores, que em sua maioria também reportam apenas ROUGE-L, ROUGE-1 e ROUGE-2.

4. Resultados

A Tabela 1 mostra os resultados da tarefa de QA usando o modelo Sabiá-7B após o FT (Fine-tuned A), e após o FT usando RAG (Fine-Tuned B). Os resultados indicam que o desempenho do modelo diminuiu consideravelmente após o FT quando combinado com o método RAG. Essa queda pode ser atribuída a dificuldades do modelo em integrar eficientemente informações recuperadas ou limitações na adaptação ao método RAG.

Tabela 1. Resultados da tarefa de QA usando o modelo Sabiá-7B após o FT (Fine-tuned A), e após o FT usando RAG (Fine-Tuned B).

	Model	**Método**	**R1**	**R2**	**RL**	**RLSum**
Fine-tuned A	Sabiá-7B	QLoRA	**0.483**	**0.329**	**0.458**	**0.457**
Fine-tuned B	Sabiá-7B	QLoRA + RAG	0.452	0.305	0.428	0.427

Como exemplo, para a pergunta: *O que a alta hidrodinâmica local provoca?*, o modelo Fine-tuned A retornou como resposta *"A mistura constante da água do estuário."*, enquanto o modelo Fine-tuned B – com RAG – retornou *"A compressão do planeta."*. A resposta referência para essa amostra é *"Constante mistura das águas do estuário."*. O modelo Fine-tuned B, nesse caso, foi influenciado pelo contexto adicional, que fala de fenômenos astronômicos, e não conseguiu desambiguar entre os contextos concorrentes.

Esse fenômeno também foi visto no trabalho de Yazan et al. [2024] que, além de demonstrar que o sucesso da integração do RAG em modelos quantizados depende de uma harmonia entre a capacidade do modelo em processar contextos recuperados e a qualidade das informações recuperadas, apontam também que o modelo LLaMA-2 demonstra uma sensibilidade maior à quantização, especialmente quando o número de documentos recuperados aumenta, o que pode dificultar seu desempenho em contextos mais longos.

Visto que a arquitetura do modelo LLaMA-2 é bastante semelhante à estrutura do LLaMA-1 [Touvron et al., 2023], o qual é base do modelo Sabiá-7B, podemos inferir que o Sabiá-7B herda muitas das características e limitações do LLaMA-1, incluindo a sensibilidade a contextos longos. Embora a ampliação do contexto possa parecer uma estratégia promissora, especialmente em modelos quantizados que utilizam o LLaMA como arquitetura subjacente, os resultados obtidos indicam que essa abordagem não alcança os efeitos desejados. Dessa forma, este trabalho complementa a análise de Yazan et al. [2024] ao fornecer evidências adicionais no domínio da língua portuguesa e em QA.

5. Conclusão

Embora o uso do RAG tenha introduzido mais contexto para a tarefa de QA, a adição das informações afetou negativamente o modelo Sabiá-7B devido à sua sensibilidade

a contextos ampliados. Este estudo contribui para a compreensão das interações entre quantização e RAG em modelos de grande escala para a língua portuguesa. Trabalhos futuros podem explorar abordagens alternativas de recuperação de informações e aplicar esses experimentos a outros modelos para ampliar o entendimento sobre a interação entre RAG e geração de texto.

6. Agradecimentos

Agradeço ao Professor Anderson Rocha Tavares pela correção deste trabalho e pelas observações construtivas que contribuíram para seu aperfeiçoamento.

Referências

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3), 2012.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*, 2019.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models. In Murilo C. Naldi and Reinaldo A. C. Bianchi, editors, *Intelligent Systems*, pages 226–240, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-45392-2.

Paulo Pirozelli, Marcos M José, Igor Silveira, Flávio Nakasato, Sarajane M Peres, Anarosa AF Brandão, Anna HR Costa, and Fabio G Cozman. Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 6(1):29–63, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Mert Yazan, Suzan Verberne, and Frederik Situmeang. The impact of quantization on retrieval-augmented generation: An analysis of small llms. 2024. URL https://arxiv.org/abs/2406.10251.