Simpósio de Integração Acadêmica

Simpósio de Integração Acadêmica

"Das Montanhas de Minas ao Oceano: Os Caminhos da Ciência para um Futuro Sustentável"

SIA UFV 2025

UFV Universidade Federal de Viçosa

Avanços no Desempenho, Privacidade e Explicabilidade para Modelos de Inteligência Artificial

Júlio César S. Oliveira e Fabrício A. Silva

Ciências Exatas e da Terra

Pesquisa

Introdução

Esse trabalho traz uma pesquisa sobre o uso da Inteligência Artificial que aumentou significativamente nos últimos anos, não apenas entre pessoas já inseridas no âmbito da Tecnologia da Informação, mas também alcançou públicos diversos, passando a ocupar vários ambientes profissionais e pessoais. Por outro lado, os grandes modelos de linguagem (LLMs), como ChatGPT ou o Gemini, ainda apresentam diversas limitações, tanto pela inviabilidade de serem treinados e utilizados com baixos custos computacionais quanto pelos desafios relacionados à privacidade dos dados.

Objetivos

O projeto tem como objetivo estudar métodos e implementações que viabilizem o uso de modelos menores (SLMs) para empresas e pessoas, de modo que tenham acesso a essa tecnologia mais acessível, segura e que atenda às suas demandas, preservando suas informações. No que diz respeito à privacidade, o uso local de modelos SLMs pré-treinados oferece maior segurança e privacidade de suas informações, uma vez que, ao utilizar uma API, especialmente de modelos closed source, como o ChatGPT, informações sensíveis podem ser coletadas e armazenadas em bancos de dados externos, que podem ser acessados e utilizados por terceiros, sem que o usuário tenha controle sobre esse processo.

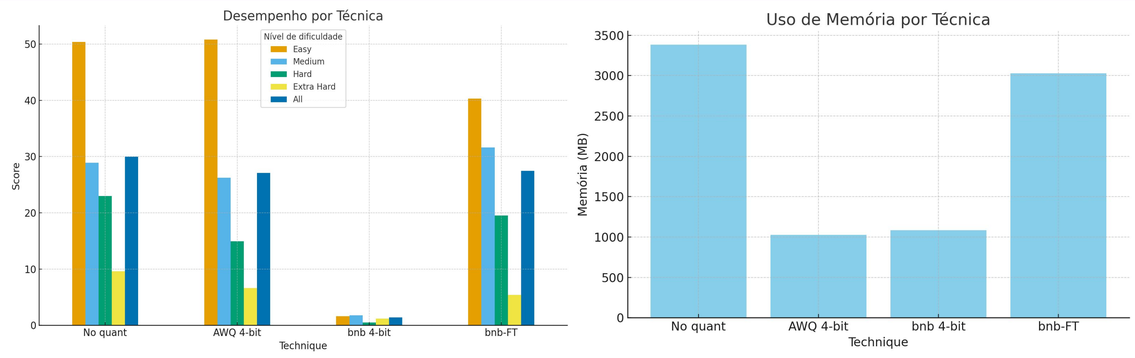
Material e Métodos ou Metodologia

Os SLMs, apesar do menor custo computacional e maior privacidade, perdem em qualidade e desempenho quando comparados às LLMs. Portanto, busca-se, neste Trabalho de Pesquisa, como melhorar a precisão e eficiência desses modelos menores, mantendo o baixo custo e a segurança dos dados. Para alcançar esse objetivo, foram testadas diferentes formas de reduzir o tamanho dos modelos por meio da Quantização e aumentar o desempenho com o Fine Tuning desses modelos. Para tanto, buscou-se na literatura científica o funcionamento das IAs por trás da camada que, usualmente, o usuário interage. Posteriormente, investigou-se formas já conhecidas de alcançar essa eficiência e comprimir os modelos. Dentre os métodos, foram experimentadas em distintos modelos diferentes técnicas de quantização, como o Quantization-Aware Training (QAT) e o Activation-Aware Weight Quantization (AWQ), além de técnicas de Fine Tuning, com o objetivo de coletar e analisar suas métricas de desempenho e custo computacional.

Apoio Financeiro

CAPES FAPEMIG CNPq

Resultados e/ou Ações Desenvolvidas



Gráficos de métricas do modelo Llama-3.2-1B-Instruct

As ações desenvolvidas tiveram como objetivo selecionar SLMs open-source de propósito geral já pré-treinados e submetê-los a diferentes testes. Os modelos escolhidos foram o TinyLlama 1.1B, o Llama 3.21B e o StableCode Instruct 3B, cujas métricas foram coletadas e analisadas por meio de distintas técnicas aplicadas sobre eles. Para cada modelo, os mesmos testes foram realizados em diferentes condições: inicialmente, avaliou-se o desempenho original, sem alterações; em seguida, aplicaram-se os métodos de quantização bnb e AWQ-4bits, a fim de verificar o impacto na performance. Por fim, cada modelo foi submetido ao Fine-tuning supervisionado, tanto com quanto sem quantização, permitindo avaliar de forma mais abrangente seu comportamento.

Conclusões

Os resultados e ações desenvolvidas buscam demonstrar que o uso de SLMs open-source representa uma alternativa viável e promissora frente às limitações de custo, privacidade e acessibilidade impostas pelos LLMs proprietários. Embora esses modelos menores apresentam desempenho inferior em alguns cenários, as técnicas de quantização e fine-tuning aplicadas mostraram-se eficazes para reduzir custos computacionais e, ao mesmo tempo, melhorar a eficiência e a precisão dos modelos testados. Assim, este trabalho evidencia que é possível alcançar um equilíbrio entre desempenho, privacidade e viabilidade de uso, abrindo caminho para soluções de Inteligência Artificial mais inclusivas e seguras, capazes de atender às necessidades tanto de indivíduos quanto de organizações em diferentes contextos.

Bibliografia

SILVA, Letícia O.; SILVA, Paulo H. C.; SILVA, Fabrício A.. Leis de Escala para Text-to-SQL: Um Estudo sobre a Relação entre Tamanho e Desempenho de Modelos de Linguagem. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 40. , 2025, Fortaleza/CE. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 140-153. ISSN 2763-8979. DOI: https://doi.org/10.5753/sbbd.2025.247042.