ComunitárIA: Viabilidade técnica e confiança do usuário na execução local de Modelos de Linguagem (LMM) Open-Source em Ambientes Comunitários

Barreto, Luciano[1]

Financiamento 04/2024/PROPPI - Fluxo Contínuo

1 Introdução

A popularização dos Modelos de Linguagem de Grande Escala (*Large Language Models - LLMs*) vem transformando atividades digitais como a geração automática de texto, o apoio à pesquisa e a síntese de informações (BOMMASANI et al., 2021). Entretanto, o predomínio de soluções proprietárias, mantidas por grandes empresas internacionais, levanta preocupações sobre privacidade, soberania informacional e dependência tecnológica (NEEL; CHANG, 2023; KPMG BRASIL, 2023).

Como alternativa, ferramentas *open-source* permite que modelos abertos sejam executados localmente em *hardware* de consumidor, com GPUs dedicadas, oferecendo maior autonomia e controle dos dados, em consonância com a Lei Geral de Proteção de Dados Pessoais - LGPD (BRASIL, 2018). Nesse contexto, este projeto buscou implantar e avaliar a viabilidade de um servidor local comunitário de LLMs no IFSC Palhoça, utilizando *frameworks* open-source e infraestrutura de hardware acessível.

2 Objetivos

O objetivo principal do estudo foi avaliar a viabilidade técnica da execução local de *LLMs* open-source em hardware de consumidor, investigando métricas de latência, uso de recursos e qualidade sintática das respostas.

3 Metodologia

O projeto foi desenvolvido em cinco etapas:

1. Revisão bibliográfica sobre LLMs, privacidade de dados e execução local (ANPD, 2023; XITE.AI, 2024).
2. Instalação e configuração dos *frameworks* em ambiente *Windows*, com registro em vídeo de tutoriais para replicabilidade.
3. Integração via API com códigos em Python para automação dos testes.
4. Execução de experimentos com prompts padronizados, registrando métricas de latência, consumo de recursos e qualidade das respostas.
5. Comparação e análise entre os diferentes *backends* e modelos, com base em medições repetidas nos dois computadores disponíveis.

4 Resultados

Foram realizadas duas análises dos resultados obtidos: uma com relação a qualidade das questões geradas por essa tupla *backend*/modelo e o tempo de resposta para a geração destas questões. Abaixo a discussão dos resultados.

Foram geradas e avaliadas aproximadamente 2.440 questões distribuídas nas disciplinas de Sociologia, História, Física, Português e Química, utilizando três modelos open-source :*mistral-7b-instruct-v0.2, huggingfaceh4/zephyr-7b-beta e meta-llama-3.1-8b-instruct-hf*, em diferentes backends (*LM Studio*, *Ollama* e *Text Generation WebUI*).

[1]Servidor Docente IFSC - PHB - luciano.barreto@ifsc.edu.br

Os testes foram executados em dois computadores: *Ryzen 7 5700X + RTX 3060 Ti* e *Ryzen 5 5600GT + RTX 4060*. Os tempos de resposta e telemetria foram registrados à parte;

Os resultados brutos foram enviados para o *LMM ChatGPT* (*GPT-5 Thinking*), que aplicou verificações automáticas e uma rubrica pedagógica heurística. Esse ambiente é adequado para a análise porque combina validadores objetivos de formato (presença de "Enunciado:", exatamente 4 alternativas A-D, "Resposta correta:" e "Justificativa:") com checagens de consistência textual (se a justificativa sustenta a alternativa marcada) e métricas agregadas por disciplina/modelo/*backend*. Em seguida, a rubrica atribuiu notas aos eixos de contextualização, clareza do problema, plausibilidade/distinção das alternativas, produzindo análises comparativas e amostras de itens para revisão humana.

Em linhas gerais, a coleção de questões geradas pelos LLMs open-source apresentou boa aderência ao template (score médio de conformidade ~90,5/100; 97,5% das questões com quatro alternativas corretas; ~86% com "Enunciado:"; ~87% com "Resposta correta:" e ~88% com "Justificativa:"), com poucas reticências (≈2%) e raras duplicações de alternativas (≈0,4%).

Variações do *LLaMA 3.1-8B* foram as mais consistentes; o *Text Generation WebUI* concentrou a maioria das quebras de formato, sugerindo ajustes de *template*/*decoding*. A comparação entre GPUs indicou diferenças pequenas na qualidade textual, geralmente explicadas por parametrização e não pelo *hardware*. Pontos fortes: estrutura "*ENEM-like*" frequente, alternativas completas e justificativas geralmente presentes.

4.1 Latência na geração das questões

Os testes realizados mostraram que, mesmo em cenários curtos de geração de questões, as latências médias ficaram quase sempre abaixo de 10 segundos por questão nos modelos *Mistral* e *Llama* rodando em *lmstudio* e *ollama*, enquanto o *Zephyr* variou mais, especialmente no *Text Generation WebUI* , mas ainda assim dentro de uma faixa aceitável (em torno de 11-16 segundos). Isso indica que todos os backends testados são viáveis para uso interativo em ambiente local, já que os tempos não comprometem a experiência de estudo ou consulta.

Um ponto interessante foi a comparação entre GPUs: a *RTX 3060* apresentou tempos consistentemente menores que a *RTX 4060*, o que provavelmente se deve à influência do processador com maior poder de processamento (*Ryzen 7 5700X*) no sistema da *RTX 3060*, reduzindo gargalos de pré-processamento e coordenação *CPU-GPU*. Isso reforça que a viabilidade do uso desses ambientes não depende apenas da *GPU* unicamente, mas também do equilíbrio da máquina como um todo. Em termos práticos, qualquer uma das combinações testadas já se mostra suficiente para suportar fluxos de estudo, elaboração de questões ou uso educacional local, sendo a escolha do *backend* e do modelo mais relevante para garantir estabilidade e previsibilidade de desempenho.

5 Conclusão

O projeto confirmou que a execução local de *LLMs open-source* em *hardware* de consumidor é não apenas tecnicamente viável, mas também relevante para o contexto educacional. A combinação de conformidade elevada no formato das questões e latências médias abaixo de 10 segundos demonstra que tais sistemas estão prontos para uso em fluxos reais de estudo, elaboração de simulados e apoio pedagógico. A análise revelou ainda que a escolha do *backend* e da configuração de execução influencia mais o desempenho do que a própria *GPU*, reforçando a importância de considerar o equilíbrio do sistema como um todo.

A pesquisa contribui para a ciência aberta e a autonomia institucional ao documentar um modelo replicável de implantação de IA local, sustentado por software livre e infraestrutura acessível. Além de reduzir a dependência de plataformas externas e riscos de

exposição de dados, o projeto abre espaço para iniciativas de ensino, pesquisa e extensão baseadas em IA comunitária, favorecendo a formação de docentes e estudantes no uso crítico e ético dessas tecnologias.

Referências

ANPD. **Nota Técnica nº 16/2023/CGTP/ANPD. Sugestões de incidência legislativa em projetos de lei sobre a regulação da Inteligência Artificial no Brasil, com foco no PL nº 2338/2023**. Autoridade Nacional de Proteção de Dados, 2023. Disponível em: https://www.gov.br/anpd/pt-br. Acesso em: 25 maio 2025.

BOMMASANI, R. et al. **On the Opportunities and Risks of Foundation Model**s. arXiv preprint arXiv:2108.07258, 2021. Disponível em: https://arxiv.org/abs/2108.07258. Acesso em: 25 maio 2025.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD)**. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 25 maio 2025.

KPMG BRASIL. **O desafio da privacidade em um mundo com crescente uso de IA**. 2023. Disponível em: https://kpmg.com/br/pt/home/insights/2023/12/desafio-privacidade-mundo-crescente-uso-ia.html. Acesso em: 25 maio 2025.

NEEL, S.; CHANG, **P. Privacy Issues in Large Language Models: A Survey**. arXiv preprint arXiv:2312.06717, 2023. Disponível em: https://arxiv.org/abs/2312.06717. Acesso em: 25 maio 2025.

XITE.AI. On-Premise LLM: **Ensuring Data Privacy & Control. 2024**. Disponível em: https://xite.ai/blogs/open-source-llm-on-premise-ensuring-data-privacy-in-the-age-of-ai/. Acesso em: 25 maio 2025.

LM STUDIO. **LM Studio: Run local LLMs**. Disponível em: https://lmstudio.ai/. Acesso em: 22 set. 2025.

OLLAMA. **Ollama - Run Llama 3, Mistral, Gemma, and other models locally**. Disponível em: https://ollama.com/. Acesso em: 22 set. 2025.

OOBABOOGA. **Text Generation WebUI**. Disponível em: https://github.com/oobabooga/text-generation-webui. Acesso em: 22 set. 2025.

HUGGING FACE. **Model Hub - Biblioteca de modelos open-source**. Disponível em: https://huggingface.co/models. Acesso em: 22 set. 2025.

OLLAMA2. **Model Library - Catálogo de modelos disponíveis para execução local.** Disponível em: https://ollama.com/library. Acesso em: 22 set. 2025.