AVALIAÇÃO COMPARATIVA DE MODELOS DE LINGUAGEM DE PEQUENO PORTE NA TRADUÇÃO PARA O PORTUGUÊS

BIANCA LOPES GONÇALVES DE MORAIS[1]
GUSTAVO HENRIQUE BRAZ FRANCO[2]
LUCAS ARAÚJO SILVA DOS REIS[3]
LUCIANO GONÇALVES DE CARVALHO[4]

RESUMO

Este estudo avalia o desempenho de dois modelos de linguagem *open source* amplamente reconhecidos em 2024: Llama 3.1 8B Instruct e Mistral Instruct 7B, com foco na tradução do inglês para o português. As traduções foram analisadas com base nas métricas BLEU e METEOR, que são baseadas em n-gramas (sequências de tokens). Os experimentos utilizaram conjuntos de dados abertos disponíveis do OPUS e foram conduzidos localmente. O Llama 3.1 8B Instruct apresentou maior consistência e melhor desempenho geral, enquanto o Mistral Instruct mostrou maior variação de desempenho. Ambos os modelos demonstraram limitações ao lidar com expressões e construções mais complexas, mas, com os resultados apresentados, há uma possibilidade de uso para determinadas atividades.

Palavras-chave: Geração de texto; Llama 3.1 8B Instruct; Mistral 7b Instruct; Modelos de linguagem; tradução português.

ABSTRACT

This study evaluates the performance of two widely recognized open-source language models in 2024: Llama 3.1 8B Instruct and Mistral Instruct 7B, focusing on the translation from English to Portuguese. The translations were analyzed based on the BLEU and METEOR metrics, which are based on n-grams (sequences of tokens). The experiments utilized publicly available open datasets from OPUS and were conducted locally. The Llama 3.1 8B Instruct demonstrated greater consistency and better overall performance, while the Mistral Instruct showed more variation in performance. Both models exhibited limitations when dealing with expressions and more complex constructions; however, given the results presented, there is potential for use in certain activities.

Keywords: Text generation; Llama 3.1 8B Instruct; Mistral 7b Instruct; Language models; Portuguese translation.

[1]Graduanda, Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Mogi das Cruzes - FATEC-MC. Mogi das Cruzes-SP. E-mail: bianca.morais4@fatec.sp.gov.br .
[2]Graduando, Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Mogi das Cruzes - FATEC-MC. Mogi das Cruzes-SP.
[3]Graduando, Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Mogi das Cruzes - FATEC-MC. Mogi das Cruzes-SP.
[4]Docente, Faculdade de Tecnologia de Mogi das Cruzes - FATEC-MC. Mogi das Cruzes-SP.

INTRODUÇÃO

Com o avanço dos Large Language Models (LLMs), como o ChatGPT, alternativa open source, como o modelo Llama da Meta, têm se destacado pela sua acessibilidade e desempenho competitivo em diversas tarefas de processamento de linguagem natural (PLN), incluindo resumo de textos, geração de conteúdo e respostas a perguntas.

Entretanto, quando usamos essa tecnologia para traduções, especialmente para o português brasileiro, enfrentamos desafios únicos devido às suas particularidades linguísticas.

A pesquisa visa fornecer insights sobre o desempenho e as limitações dos modelos open source na tradução para o português, ajudando na escolha e implementação de ferramentas de tradução mais eficazes e acessíveis. O resultado é particularmente relevante para desenvolvedores e pesquisadores que necessitam de modelos de linguagem capazes de operar em ambientes com recursos computacionais limitados e que buscam alternativas aos serviços comerciais, muitas vezes inacessíveis para pequenos grupos ou empresas.

MATERIAL E MÉTODOS

Os experimentos foram conduzidos em um computador local equipado com uma placa de vídeo NVIDIA GeForce RTX 2060 SUPER com 8 GB de memória dedicada, utilizando a linguagem de programação Python 3.10. Os modelos de linguagem Llama3.1 8b instruct e Mistral Instruct 7B, ambos de código aberto, foram obtidos do repositório Hugging Face e executados localmente.

Para a avaliação da qualidade das traduções, foram utilizadas as métricas BLEU e METEOR, disponíveis no kit de ferramentas NLTK (Natural Language Toolkit), amplamente reconhecidas na área de processamento de linguagem natural.

Conjunto de Dados

Os textos utilizados neste estudo foram extraídos da coleção OPUS, uma ampla base de dados de textos traduzidos disponível na web. Dentro do OPUS,

foram selecionados dois conjuntos específicos: "Books" (sentenças com 1.404 e 25.407 tokens), que inclui livros de domínio público, e "MDN Web Docs" (sentenças com 12.666 e 134.557 tokens), um corpus bilíngue (inglês-português) que é um projeto colaborativo de código aberto que documenta tecnologias da web.

Comunicação com os modelos

Para a comunicação com os modelos, foram utilizadas as seguintes instruções: para Mistral, o input foi “Translate the following text to Brazilian Portuguese. Provide only the translation with no additional notes, explanations, or formatting. Exclude any extra information like 'Note:' or similar. Just the translation.”. Já para o modelo Llama 3.1 8B, o comando foi “Translate to Brazilian Portuguese. Objective response. Exclude notes. Don't repeat the input.” Essa abordagem garantiu que as traduções fossem diretas e sem informações adicionais, permitindo uma comparação mais clara entre os resultados dos modelos.

REFERENCIAL TEÓRICO

Modelos de linguagem open source, como o Llama 3.1 8B Instruct da Meta e o Mistral Instruct 7B da Mistral AI, são alguns dos mais populares e grandemente utilizados na atualidade. Portanto, é razoável supor que os resultados obtidos neste estudo reflitam o desempenho geral dessa classe de modelos. Embora outros modelos open source possam apresentar variações em suas arquiteturas e dados de treinamento, é provável que suas capacidades na tradução automática para o português brasileiro sejam comparáveis às observadas neste estudo.

INSTRUCT (Instruções e Aplicações)

Os modelos *instruct* são desenvolvidos com o objetivo de melhorar a capacidade dos sistemas de inteligência artificial em interpretar e seguir instruções de maneira mais eficaz. Diferentemente dos modelos de linguagem gerais, que podem gerar respostas mais vagas ou amplas, os modelos *instruct* são treinados especificamente para executar tarefas com base em comandos explícitos fornecidos pelos usuários. Esses modelos são mais adequados para aplicações onde a

precisão e a aderência às instruções são essenciais, como tradução automática, geração de código e execução de tarefas sequenciais complexas

BLEU - Bilingual Evaluation Understudy

A métrica BLEU avalia traduções comparando a tradução gerada com uma ou mais traduções de referência, analisando a coincidência de n-gramas (sequências de n palavras). O BLEU-1 considera unigramas, o BLEU-2 considera bigramas, o BLEU-3 considera trigramas e o BLEU-4 considera quatro-gramas. Cada nível adiciona mais contexto à avaliação, e o BLEU-4 é o mais usado porque oferece a melhor combinação de precisão e fluidez ao capturar sequências mais longas, proporcionando uma avaliação mais completa da qualidade da tradução.

Kishore Papineni, cientista da computação e pesquisador da IBM, é um dos principais desenvolvedores do BLEU, que se tornou uma métrica padrão para avaliar traduções automáticas. No trabalho *'BLEU: A Method for Automatic Evaluation of Machine Translation'* (2002), ele e seus co-autores explicam que o BLEU foi criado para quantificar a qualidade das traduções de forma objetiva, utilizando uma abordagem estatística (Papineni et al., 2002)."

METEOR - Metric For Evaluation Of Translation With Explicit Ordering

A métrica METEOR também compara a tradução gerada com uma ou mais traduções de referência, mas leva em consideração não apenas a coincidência de n-gramas, mas também a ordem das palavras na frase e a correspondência de palavras-chave (stemming). Além disso, a métrica METEOR utiliza sinônimos e paráfrases para aumentar a correspondência entre as traduções.

Alon Lavie, professor de Ciência da Computação e Linguística Computacional, coautor do artigo 'An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments' (2005), aponta que o METEOR foi projetado para corrigir as limitações do BLEU, incorporando elementos semânticos e sinônimos para uma avaliação mais rica (Banerjee & Lavie, 2005).

RESULTADOS E DISCUSSÃO

A qualidade e a quantidade dos dados de treinamento são fatores cruciais no desempenho de modelos de linguagem de grande porte (LLMs) em tarefas de tradução automática. A escassez de dados de alta qualidade em determinados idiomas pode resultar em traduções inconsistentes e imprecisas (Kaplan et al., 2020).

Desempenho Llama 3.1 Dataset Books

No Quadro 1, observa-se que o desempenho geral do modelo de tradução deixou a desejar. Em diversos casos, a ferramenta apresentou resultados com pontuação zero. Contudo, houve exemplos notáveis de traduções bem-sucedidas. Um desses casos pode ser visto na frase: Original: "Você devia ter terminado", disse o Rei. Tradução: "Você devia ter terminado", disse o Rei. Aqui, a tradução manteve a integridade da frase original, inclusive preservando as aspas e a estrutura fiel ao inglês: "You ought to have finished," said the King. No entanto, também foram registrados casos de erros graves. Por exemplo, na frase: Original: "A sua primeira ideia era a de que, de algum modo, ela tinha caído no mar. E, nesse caso, posso voltar de trem", falou para si mesma. Tradução: "O seu primeiro pensamento foi que ela teria caído no mar, e no caso disso, posso voltar de trem", disse a si mesma. Nota-se aqui uma falta de concordância adequada em “no caso disso”, o que compromete a fluidez e a naturalidade da tradução.

Quadro 1. Resultado tradução Books.

Métricas	Bleu - 1	Bleu - 2	Bleu - 3	Bleu - 4	Meteor
Nota Llama 3.1 8B Q4	0.39	0.28	0.21	0.158	0.39

Fonte: Elaborado pelos autores. (2024).

Desempenho Llama 3.1 7B Web docs

Os resultados obtidos com o dataset MDN Web Docs estão apresentados no Quadro 2. Observa-se um aumento significativo nas métricas finais, além de

exemplos que demonstram uma boa capacidade de tradução técnica. Um exemplo notável é a tradução da frase em inglês: "A function is a code snippet that can be called by other code or by itself, or a {{Glossary(“variable”)}} that refers to the function." para "Uma função é um fragmento de código que pode ser invocado por outro código, por si mesma, ou uma {{Glossary(“variável”)}} que se refere à função." Essa tradução, de certa forma, captura adequadamente termos técnicos essenciais.

Entretanto, algumas falhas ainda são evidentes. Por exemplo o erro encontado na seguinte tradução, em vez de "Este é a vasta maioria", a forma correta seria "Esta é a vasta maioria", considerando que "maioria" é um substantivo feminino. Além disso, o uso de "código" no singular não reflete com precisão a intenção da frase original, que se refere a "maioria dos códigos", inglês original: “This is the vast majority of JavaScript code on the Web”.

Quadro 2. Resultado tradução Web DOCS.

Métricas	Bleu - 1	Bleu - 2	Bleu - 3	Bleu - 4	Meteor
Nota Llama 3.1 8B Q4	0.43	0.35	0.29	0.24	0.42

Fonte: Elaborado pelos autores, (2024).

Desempenho Mistral Books

Como demonstrado pelas métricas no Quadro 3, o Mistral apresentou diversos problemas relacionados à qualidade de suas traduções, resultando em uma pontuação inferior em comparação ao Llama. Um exemplo claro disso pode ser observado na tradução da frase em inglês: "How doth the little crocodile." No original em português, essa frase foi corretamente traduzida como: "Como o pequeno crocodilo." No entanto, a tradução realizada pelo Mistral gerou: "Como faz a pequena crocodila." Nesse caso, o Mistral comete dois erros: primeiro, ao utilizar "crocodila", uma forma inexistente na língua portuguesa, revelando dificuldade em identificar o gênero correto; segundo, ao alterar a estrutura da frase, substituindo "Como" por "Como faz", o que resulta em uma tradução menos direta e fluida. Embora o Mistral tente preservar o sentido geral da frase, ele não alcança o nível de precisão e naturalidade esperado, comprometendo significativamente a qualidade da tradução.

Quadro 3. Resultado tradução Books.

Métricas	Bleu - 1	Bleu - 2	Bleu - 3	Bleu - 4	Meteor
Nota MISTRAL	0.38	0.27	0.20	0.14	0.39

Fonte: Elaborado pelos autores. (2024).

Desempenho Mistral Web docs

No Quadro 4, são apresentados os resultados obtidos com o dataset Web Docs, nos quais o modelo obteve notas piores em relação às obtidas no dataset Books. Esses resultados oferecem uma visão mais detalhada sobre a capacidade do modelo Mistral em lidar com textos técnicos. Entretanto, mesmo com resultados inferiores, o Mistral consegue um desempenho satisfatório para determinadas situações. Um exemplo é na tradução da frase em inglês: "JavaScript (or C/C++ using Emscripten to compile to JavaScript)". A tradução realizada pelo modelo manteve-se fiel ao original, resultando em: "JavaScript (ou C/C++ usando Emscripten para compilar para JavaScript)". Isso indica que, apesar das limitações observadas, o Mistral tem certa capacidade de manter a equivalência semântica em frases técnicas um pouco mais complexas.

Quadro 4. Resultado tradução Web DOCS

Métricas	Bleu - 1	Bleu - 2	Bleu - 3	Bleu - 4	Meteor
Nota MISTRAL	0.29	0.22	0.17	0.14	0.34

Fonte: Elaborado pelos autores, (2024).

Alucinação

Durante os experimentos, mesmo utilizando modelos voltados para instruções específicas (instruct models), observamos casos de alucinação — ou seja, situações em que o modelo gera traduções não relacionadas ao conteúdo original. Um exemplo marcante foi a tradução do termo "GamesSidebar", em que o modelo Mistral produziu uma lista de categorias de jogos, fugindo completamente do esperado. Esse tipo de comportamento compromete a qualidade de tradução em textos técnicos e exige atenção redobrada na validação dos resultados.

Esse fenômeno de alucinação é particularmente prejudicial em contextos onde a precisão terminológica é crucial, como na tradução de documentação técnica, contratos, ou material legal. Nesses casos, uma tradução incorreta ou inventada pode resultar em mal-entendidos graves ou até mesmo consequências legais. Pesquisas apontam que, embora os modelos de linguagem avancem na geração de texto, eles frequentemente falham na compreensão semântica, resultando em alucinações e gerações incoerentes, como observado por Bender e Koller (2020) ao discutir os limites da compreensão dos modelos de linguagem natural.

Aplicação no dia a dia

Modelos de geração de texto, como os avaliados, demonstram um potencial para as tarefas cotidianas, desde auxiliar em pesquisas até lidar com documentos técnicos ou conversas informais. Sua versatilidade pode ser especialmente valiosa em situações que exigem uma comunicação rápida e eficaz entre idiomas.

No entanto, é importante que os usuários estejam atentos às variações na qualidade das traduções, principalmente em contextos que exigem alta precisão e compreensão de nuances, como em traduções técnicas ou jurídicas. Embora ambos os modelos possam atender a demandas simples e curtas, inconsistências—particularmente com expressões idiomáticas e construções sintáticas complexas—podem comprometer a fluidez e naturalidade do texto traduzido, como visto nos parágrafos anteriores.

Um ponto positivo significativo desses modelos menores e de código aberto é a possibilidade de executá-los localmente, o que permite que usuários com recursos computacionais limitados os utilizem em seus próprios dispositivos. Isso democratiza o acesso a ferramentas avançadas de tradução e torna essas soluções mais acessíveis para indivíduos ou pequenas organizações que não têm a infraestrutura necessária para modelos maiores e comerciais.

Esses modelos oferecem soluções práticas para necessidades comuns de tradução, mas exigem cuidado ao serem usados. Com o aprimoramento contínuo, especialmente na melhoria dos dados de treinamento para línguas específicas como o português em modelos menores, essas ferramentas podem se tornar ainda mais

confiáveis para o uso robusto. De acordo com Vaswani et al. (2017), "a atenção é um mecanismo que permite que os modelos foquem em partes específicas da entrada, melhorando a eficiência e a eficácia da tradução automática".

CONCLUSÃO

Este estudo apresentou uma avaliação comparativa entre dois modelos de linguagem de pequeno porte, Llama 3.1 8B Instruct e Mistral Instruct 7B, no contexto da tradução automática para o português. Os resultados obtidos, medidos pelas métricas BLEU e METEOR, evidenciaram diferenças no desempenho dos modelos.

O Llama demonstrou certa consistência, com pontuações finais mais altas em todas as métricas avaliadas nos dois conjuntos de dados presentes no trabalho, possivelmente devido ao treinamento em um conjunto de dados mais abrangente, incluindo textos em português. Já o Mistral, apesar de apresentar resultados inferiores, alguns de seus exemplos se destacaram, trazendo limitações, mas mostrando que não está tão distante.

Embora ambos os modelos apresentem restrições, especialmente ao lidar com expressões e construções sintáticas complexas, os resultados indicam que seu uso é viável, ainda que com certas ressalvas. As discrepâncias observadas destacam a necessidade contínua de aprimoramento dos modelos e dos conjuntos de dados para assegurar uma tradução automática mais precisa e fluida.

Essa avaliação comparativa contribui para uma melhor compreensão das capacidades e limitações dos modelos de código aberto na tradução para o português, oferecendo insights relevantes para desenvolvedores e pesquisadores que buscam soluções tecnológicas acessíveis e eficientes.

REFERÊNCIAS BIBLIOGRÁFICAS

BANERJEE, S.; LAVIE, A. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. Disponível em: <https://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf>. Acesso em: 22 set. 2024.

BENDER, E. M.; KOLLER, A. **Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data**. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.463.pdf>. Acesso em: 12 set. 2024.

KAPLAN, J. et al. **Scaling Laws for Neural Language Models**. 2020. Disponível em: <https://arxiv.org/pdf/2001.08361>. Acesso em: 12 set. 2024.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W. **BLEU: a Method for Automatic Evaluation of Machine Translation**. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: AssociatiVon for Computational Linguistics, 2002. Disponível em: <https://aclanthology.org/P02-1040.pdf>. Acesso em: 20 set. 2024.

SAADANY, H.; ORASAN, C. **BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text**. Disponível em: <http://arxiv.org/abs/2109.14250>. Acesso em: 17 maio 2024.

TIEDEMANN, J. **Parallel Data, Tools, and Interfaces in OPUS**. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Istanbul: European Language Resources Association, 2012. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf>. Acesso em: 6 set. 2024.

VASWANI, A. et al. **Attention is all you need**. Disponível em: <http://arxiv.org/abs/1706.03762>. Acesso em: 1 jun. 2024.