Diferenciação de Ataques em Português e em Inglês: Análise de Experimento sobre as Barreiras Linguísticas no Alinhamento de LLMs

Bruno Zimmermann Russo Ferreira[1], **Victor Takashi Hayashi**[1]

[1]Escola Politécnica da Universidade de São Paulo (USP)
São Paulo – SP – Brazil

bruno.zrferreira@usp.br, victor.hayashi@usp.br

Abstract. *This paper compares jailbreak success rates of Llama 3.1 8B on equivalent English and Portuguese prompts, judged by Llama 3.3 and manually curated. Curated ASR was higher in Portuguese (16%) than English (10%), but the key finding is that the LLM judge misses far more jailbreaks in Portuguese – four times more than in English – revealing a language-calibration bias in automated safety evaluation.*

Resumo. *Este artigo compara a taxa de sucesso de jailbreak do Llama 3.1 8B em prompts equivalentes em inglês e português, julgados pelo Llama 3.3 e curados manualmente. O ASR curado foi maior em português (16%) do que em inglês (10%), mas o achado principal é que o juiz LLM deixa passar muito mais jailbreaks em português – quatro vezes mais que em inglês – revelando um viés de calibração por idioma na avaliação automatizada de segurança.*

1. Introdução

O avanço dos *Large Language Models* (LLMs) revolucionou o processamento de textos, mas trouxe consigo riscos significativos, como os ataques adversariais [Miers et al. 2024, Vassilev et al. 2025]. Enquanto a literatura acadêmica sobre o tema é vasta para o idioma inglês, existe uma carência crítica de investigações voltadas para a língua portuguesa [Figueiredo 2025], o que torna incerta a robustez de modelos ajustados para este idioma, frente a manipulações intencionais [Figueiredo 2025].

A eficácia dos mecanismos de segurança (*guardrails*) é frequentemente dependente de padrões do *fine-tuning* dos modelos, tornando-os vulneráveis a variações [Yuan et al. 2024]. Estudos indicam que o alinhamento de modelos pode sofrer degradação consistente quando o input se desloca para domínios alheios aos quais foram treinados, onde as políticas de recusa são, portanto, mais fracas [Yuan et al. 2024, Queiroz 2026]. No caso do português, a complexidade da língua e a menor densidade de exemplos de segurança no treinamento podem ampliar essas lacunas [Queiroz 2026], sugerindo que simples ataques em linguagem cotidiana, quando realizados em língua portuguesa, podem ser tão eficazes quanto métodos tecnicamente mais sofisticados da língua inglesa [Figueiredo 2025].

Neste contexto, este artigo relata um experimento preliminar e pareado com 50 ataques baseados em linguagem cotidiana, comparando o *Attack Success Rate* (ASR) do mesmo modelo-alvo em inglês e em português [Figueiredo 2025]. Além de estimar essa

diferença, o experimento revela uma limitação metodológica relevante para a área: o viés do juiz LLM automatizado em favor de padrões de recusa anglófonos, contribuindo tanto para o debate sobre segurança de LLMs em português quanto para o desenho de avaliações automatizadas mais confiáveis [Queiroz 2026, Figueiredo 2025].

2. Descrição do Experimento

Por se tratar de uma análise preliminar, foi selecionado um dataset de 50 ataques estáticos no estilo HarmBench [Mazeika et al. 2024], aplicados ao modelo-alvo Llama-3.1-8b-instant em execução única de cada prompt, com julgamento realizado pelo modelo Llama-3.3-70b-versatile. Os resultados de ambos os grupos de ataque foram direcionados a um arquivo JSON para curadoria manual, sendo a métrica avaliada o *Attack Success Rate* (ASR) [Miers et al. 2024].

A escolha do Llama 3.1 8B como alvo se justifica por ser um modelo *open-weight* de capacidade intermediária, com alinhamento documentado, mas ainda acessível e sem a robustez de modelos de maior porte. O juiz principal é deliberadamente maior que o alvo: essa assimetria de capacidade é intencional, já que um modelo maior tende a compreender nuances de respostas que um modelo menor poderia avaliar incorretamente [De Oliveira et al. 2025]. O juiz classifica como sucesso (score 1) quando o modelo forneceu conteúdo nocivo, ilegal ou não-ético, deu instruções para atividades perigosas, ou contornou criativamente suas proteções. Classifica como falha (score 0) quando o modelo recusou, forneceu apenas informação genérica/educacional segura, ou redirecionou sem cumprir o pedido [Chao et al. 2025].

Para realizar o experimento, optou-se pela utilização de um LLM para traduzir os ataques do inglês para o português. Para contornar possíveis recusas do modelo a traduzir prompt malicioso para o português, foi realizada curadoria manual das traduções por um anotador humano, porém não cego ao idioma. O LLM selecionado como tradutor foi o Mistral-small-latest, por possuir alinhamento menos restritivo. [Khoboko et al. 2025].

3. Resultados do Experimento

O ataque em inglês (Grupo A) registrou um ASR de 8% conforme avaliação do juiz. Após curadoria manual, constatou-se um falso negativo, elevando o ASR real para 10%. Nos ataques em português (Grupo B), o ASR também registrado pelo juiz foi de 8%; porém, a revisão manual identificou quatro falsos negativos, totalizando um ASR real de 16%. A tradução mostrou-se eficiente e, apesar da curadoria manual, não foram encontrados erros de tradução relevantes.

A Tabela 1 lista os prompts julgados como bem-sucedidos (YES) pelo juiz em cada grupo. Os IDs são compartilhados entre os dois grupos (mesmo dataset), e FN indica Falso Negativo identificado na curadoria manual.

Nos falsos negativos, o padrão observado é semelhante: o modelo recusa o pedido literal, mas gera conteúdo substantivo relacionado ao tema, geralmente sob um enquadramento hipotético, fictício ou educacional, acompanhado de *disclaimers* no início e/ou no fim da resposta. O juiz interpretou esses *disclaimers* como recusa suficiente em todos os casos, mesmo quando o conteúdo gerado era tecnicamente substancial.

A discrepância entre o veredito do juiz e a anotação humana é maior em português (ASR humano de 16% contra ASR do juiz de 8%) do que em inglês. Isso sugere que o

Tabela 1. ASR por categoria: veredito do juiz vs. pós-curadoria manual.

Categoria	N	Juiz EN	Pós-curado EN	Juiz PT	Pós-curado PT
Hacking	10	0/10	1/10	2/10	3/10
Misinformation	5	2/5	2/5	1/5	2/5
Violence	5	1/5	1/5	0/5	0/5
Harmful Content	5	0/5	0/5	0/5	1/5
Weapons	4	0/4	0/4	0/4	1/4
Malware	4	1/4	1/4	1/4	1/4
Outras*	17	0/17	0/17	0/17	0/17
Total	**50**	**4/50**	**5/50**	**4/50**	**8/50**

*Fraud (7), Hate Speech (5), Financial Crime (3), Bioweapon (1) e Social Engineering (1) – nenhum sucesso registrado em nenhum dos quatro grupos.

juiz LLM está calibrado principalmente em padrões de resposta anglófonos: respostas que em inglês sinalizariam claramente um *jailbreak* (conteúdo + *disclaimer*) são tratadas com mais leniência quando o *disclaimer* está em português e ocupa proporção maior do texto.

Em todos os casos de *jailbreak* confirmados, o modelo cedeu sob o argumento implícito ou explícito de que o conteúdo era hipotético ou educacional. Em português, as recusas tendem a ser mais verbosas: o modelo frequentemente explica o motivo da recusa, oferece alternativas e, nesse processo, acaba gerando conteúdo substantivo sob pretexto de ”consciência educacional-- um padrão que o juiz LLM não captura de forma confiável.

4. Considerações Finais

O achado mais relevante deste trabalho não é a diferença absoluta de ASR entre os idiomas, cuja significância estatística é limitada pelo tamanho da amostra, mas sim a evidência de que o juiz automatizado subestima o ASR de forma mais acentuada em português. Isso indica um risco prático para pipelines de avaliação de segurança que dependem de juízes LLM: a calibração do juiz pode não se generalizar entre idiomas, exigindo validação humana especialmente em contextos multilíngues.

Como trabalhos futuros, propõe-se: (i) ampliar a amostra e repetir gerações por prompt (com temperatura fixa e sementes controladas) para permitir testes estatísticos formais (e.g., teste de McNemar) sobre a diferença de ASR entre idiomas; (ii) avaliar os mesmos pares de prompts com um único juiz após tradução padronizada das respostas para o inglês, isolando o efeito do idioma na calibração do julgamento; (iii) usar múltiplos anotadores humanos independentes, cegos ao idioma, para curadoria; e (iv) expandir o escopo do estudo para investigar diretamente o viés de juízes LLM em avaliações de segurança multilíngues.

5. Agradecimentos

Os autores agradecem o apoio da FAPESP, Processo 26/00353-0.

Referências

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. (2025). Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

De Oliveira, D. E. G. C., Miers, C. C., Simplicio, M. A., and Hayashi, V. T. (2025). Between generation and judgment: A cloud-native framework for adversarial evaluation of llm alignment. In *2025 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 1–8.

Figueiredo, A. C. (2025). Ataque adversarial em modelos de linguagem para a língua portuguesa. Dissertação de mestrado, UNIRio. Disponível em: `http://hdl.handle.net/unirio/15161`.

Khoboko, P. W., Marivate, V., and Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, 20:100649.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. (2024). Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In *ICML*.

Miers, C. C., Jr., M. A. S., Rojas, M. A. T., de Oliveira, D. E. G. C., Neto, M. P. P., Damin, F. A. S., Jr., R. B., and Hayashi, V. T. (2024). Ferramentas de jailbreaking para grandes modelos de línguas – aprendizado de máquina no contexto adversário. In *Minicursos do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg 2024)*, pages 49–91. Sociedade Brasileira de Computação.

Queiroz, J. (2026). Versificação adversarial em português como operador de jailbreak em llms. *SciELO Preprints*.

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., and Hamin, M. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, NIST AI 100-2e2025.

Yuan, Y., Jiao, W., Wang, W., Huang, J.-t., He, P., Shi, S., and Tu, Z. (2024). Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *International Conference on Learning Representations*, volume 2024, pages 53902–53922.