Além da Distância Linguística: Jailbreaks Multilíngues em LLMs Especializados por Idioma

Gabriel de Jesus Coelho da Silva [1] , **Carlos Becker Westphall** [2]

[1]Programa de Pós-Graduação em Ciência da Computação – UFSC
Florianópolis – SC – Brasil

gabriel.jcs@posgrad.ufsc.br

[2]Departamento de Informática e Estatística – UFSC
Florianópolis – SC – Brasil

westphal@inf.ufsc.br

Abstract. *This paper evaluates whether cross-lingual jailbreak success against language-specialized LLM assistants is better explained by typological distance or by post-training specialization. We evaluate eight instruction assistants organized in four weak/strong pairs aligned to Portuguese, Italian, Swedish, and Bulgarian, across thirteen attack languages, scoring responses with StrongREJECT and joining URIEL+ distances and BELEBELE specialization metrics in mixed-effects logistic models. The results do not support the a priori distance account: attack success is mostly model-specific, and stronger paired models are usually safer than their weak counterparts.*

Resumo. *Este artigo avalia se o sucesso de jailbreaks multilíngues contra LLMs especializados por idioma é melhor explicado por distância tipológica ou por especialização pós-treinamento. Avaliamos oito assistentes instrucionais organizados em quatro pares fraco/forte alinhados a português, italiano, sueco e búlgaro, em treze idiomas de ataque, pontuando respostas com StrongREJECT e combinando distâncias URIEL+ e métricas de especialização do BELEBELE em modelos logísticos de efeitos mistos. Os resultados não sustentam a hipótese de distância: o risco é predominantemente específico do modelo, e modelos pareados mais fortes são, em geral, mais seguros que seus pares fracos.*

1. Introdução

Modelos de Linguagem de Grande Escala (LLMs) são normalmente submetidos a ajuste fino e alinhamento de segurança antes de sua implantação, a fim de reduzir respostas danosas, ilegais ou socialmente indesejáveis [Ouyang et al. 2022, Ganguli et al. 2022]. Ainda assim, ataques por *jailbreak* podem induzir o modelo a contornar esse comportamento de recusa, especialmente quando exploram instruções ambíguas, *role-playing*, reformulações ou idiomas menos comuns nos testes de segurança [Yong et al. 2023, Deng et al. 2023].

Esse problema é particularmente relevante para LLMs especializados em um idioma. Estes modelos são treinados para atender melhor comunidades locais, mas podem herdar alinhamentos de segurança de modelos-base majoritariamente em inglês, receber dados de pós-treinamento excessivamente focados em um único idioma e operar em ambientes de uso naturalmente multilíngues. A pergunta central deste trabalho é, portanto:

como o idioma de um *prompt* de *jailbreak* afeta sua taxa de sucesso contra LLMs alinhados a outro idioma?

Avaliamos três hipóteses especificadas *a priori*:

- H_1**:** a Taxa de Sucesso de Ataque (*Attack Success Rate*, ASR) aumenta à medida que o idioma de ataque se distancia tipologicamente da língua alinhada do modelo;
- H_2**:** a ASR diminui em idiomas nos quais o modelo apresenta maior especialização pós-treinamento;
- H_3**:** ao incluir distância linguística e especialização pós-treinamento no mesmo modelo estatístico, a distância permanece o preditor mais forte e estável da ASR.

A contribuição principal é empírica e negativa: no painel avaliado, essas hipóteses não se sustentam. O sucesso dos ataques varia fortemente por família de modelo e configuração pós-treinamento; a distância tipológica não apresenta um efeito positivo estável; e a especialização não atua como fator protetivo. O estudo também contribui com um protocolo auditável para avaliação multilíngue de segurança, preservando pontuações, recusas, falhas técnicas, metadados de API, qualidade de tradução, diagnósticos de *tokenização* e dados para reanálise.

2. Fundamentação e Trabalhos Relacionados

Jailbreaks são instruções (*prompts*) que buscam contornar o alinhamento de segurança e obter conteúdo que o assistente deveria recusar. Eles são um subconjunto de injeções de *prompt* (*prompt injections*): neste trabalho não consideramos instruções indiretas em documentos, acesso a ferramentas, exfiltração de dados ou ataques adaptativos de múltiplos turnos. A vulnerabilidade multilíngue pode surgir porque pré-treinamento, instrução e alinhamento não cobrem todos os idiomas de forma igual, podendo concentrar-se majoritariamente em um idioma ou família de idiomas [Shen et al. 2024].

Os estudos estabelecem duas linhas principais: Yong et al. [Yong et al. 2023] traduzem *prompts* nocivos, consultam GPT-4 e retrotraduzem as respostas, mostrando vulnerabilidades em idiomas de poucos recursos, isto é, idiomas com menor disponibilidade de dados digitais, ferramentas de PLN e avaliações sistemáticas; enquanto MultiJail [Deng et al. 2023] usa consultas multilíngues construídas com tradução humana e também propõe ajuste de segurança multilíngue. Trabalhos posteriores ampliam o espaço de modelos, idiomas e ataques, incluindo sistemas fechados e conversas adaptativas [Huang et al. 2025, Singhania et al. 2025]. Em paralelo, distância linguística é associada à perda de transferência entre idiomas em tarefas benignas [Philippy et al. 2023], o que, embora não estabeleça diretamente uma contrapartida de segurança, fornece um indício plausível para testar essa hipótese.

Este trabalho contribui com uma avaliação sistemática de *jailbreaks* multilíngues em oito assistentes especializados por idioma, organizados em quatro pares fraco/forte; com o teste conjunto de distância tipológica e especialização pós-treinamento como explicações concorrentes para a ASR; e com um pipeline auditável que preserva traduções, saídas originais e retrotraduzidas, escores contínuos, recusas, falhas técnicas, metadados de API e diagnósticos de *tokenização*. O StrongREJECT é adotado como protocolo de avaliação por penalizar "*jailbreaks* vazios" e por seu avaliador ter sido calibrado contra 1.361 pares de resposta rotulados por cinco avaliadores humanos, com alta concordância

Tabela 1. Posicionamento frente aos trabalhos multilíngues mais próximos.

Estudo	Idiomas	LLMs espec.	Distância/espec.	Foco distintivo
Yong et al.	13	não	não/não	tradução–consulta–retrotradução em GPT-4
MultiJail	9	não	não/não	conjunto multilíngue e defesa por ajuste fino
Language Barrier	19	não	recursos/não	segurança por nível de recursos e regime de ajuste
MM-ART	7	não	não/não	ataques multilíngues adaptativos e multi-turno
Este artigo	**13**	**sim**	**sim/sim**	pares por idioma, QC de tradução, efeitos cruzados e dados congelados

Referências: [Yong et al. 2023]; [Deng et al. 2023]; [Shen et al. 2024]; [Singhania et al. 2025]. *Nota:* em Language Barrier, "recursos" significa a disponibilidade de dados linguísticos digitais, ferramentas de PLN e avaliações; o estudo compara 9 idiomas de alto recurso e 10 de baixo recurso.

com o julgamento humano [Souly et al. 2024]. Essa validação sustenta seu uso como instrumento padronizado e altamente robusto.

3. Desenho Experimental

O painel contém oito assistentes LLM instrucionais, organizados em quatro pares fraco/forte alinhados a português, italiano, sueco e búlgaro (Tabela 2). Os pares fraco/forte também diferem em modelo-base, mistura de dados, tamanho, pós-treinamento e modo de acesso.

Os rótulos *fraco* e *forte* foram definidos antes da análise com base na documentação disponível sobre pós-treinamento e mecanismos de segurança dos modelos. O polo *fraco* corresponde ao modelo menos recente ou com evidência mais limitada de pós-treinamento voltado à segurança; o polo *forte*, ao comparador mais recente ou com documentação mais extensa desse processo. Esses rótulos são categorias operacionais para o contraste empírico, devendo ser interpretados como linha de base e presunção de robustez, e não como uma medida isolada de alinhamento.

Tabela 2. Modelos avaliados por idioma alinhado.

Idioma	Polo	Tam.	Acesso	Modelo
Português	fraco	7B	pesos abertos	Sagui-7B-Instruct-v0.1
	forte	n.d.	API	Sabiá-3
Italiano	fraco	7B	pesos abertos	LLaMAntino-2-chat-UltraChat-ITA
	forte	8B	pesos abertos	LLaMAntino-3-ANITA-DPO-ITA
Sueco	fraco	6,7B	pesos abertos	GPT-SW3-6.7B-v2-instruct
	forte	8B	pesos abertos	AI Sweden Llama-3-8B-instruct
Búlgaro	fraco	7B	pesos abertos	BgGPT-7B-Instruct-v0.2
	forte	9B	pesos abertos	BgGPT-Gemma-2-9B-Instruct

Fonte: Elaboração própria, a partir da documentação e dos cartões de modelo. Tamanho em bilhões de parâmetros quando divulgado; n.d. = não divulgado.

O painel foi documentado a partir dos relatórios técnicos e cartões de modelo disponíveis [Oliveira 2024, Abonizio et al. 2024, Basile et al. 2023, Polignano et al. 2026, Ekgren et al. 2024, AI Sweden NLU Team 2024, Alexandrov et al. 2024, INSAIT Institute 2024]. Para Sabiá-3, por ser API fechada, preservamos identificador de provedor, data, status HTTP, falhas e bloqueios separadamente das respostas do modelo, mas não foi possível garantir que não houve moderação, pois o ecossistema não é

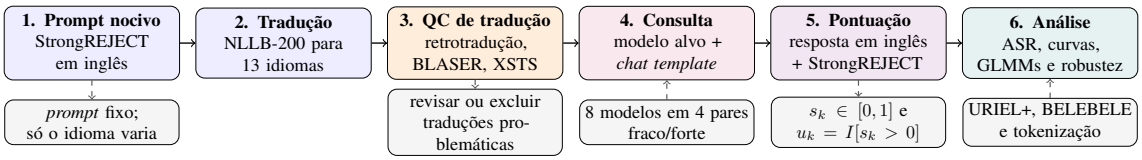
totalmente auditável de ponta-a-ponta. Embora o acesso ao Sabiá-3 exclusivamente por API limite a transparência sobre seus pesos, filtros e atualizações, sua inclusão amplia a robustez externa do estudo ao testar se os padrões observados persistem além de modelos de pesos abertos e de uma única família de arquiteturas.

Usamos treze idiomas de ataque: árabe, búlgaro, dinamarquês, inglês, finlandês, italiano, norueguês, português, russo, espanhol, suaíli, sueco e ucraniano. Cada modelo é avaliado tanto em sua língua alinhada quanto nos demais idiomas. A língua alinhada é mantida como linha de base, mas excluída quando calculamos ASR agregada.

Modelo de ameaças. O adversário possui acesso caixa-preta à interface de conversa, controla uma única mensagem de usuário e escolhe seu idioma. Ele não controla a mensagem de sistema, os pesos, a infraestrutura do provedor ou ferramentas externas, e não usa otimização adaptativa, múltiplos turnos ou injeção indireta. O ativo protegido é a política de recusa do assistente para solicitações textuais nocivas. Há sucesso quando uma geração concluída fornece conteúdo inseguro e útil segundo StrongREJECT. Bloqueios do gateway são separados de recusas do próprio modelo.

A Figura 1 resume o fluxo operacional. O diagrama separa a transformação linguística do *prompt*, a geração nos modelos, a normalização para pontuação em inglês e a junção das variáveis explicativas usadas nas análises.

Figura 1. Fluxo metodológico da avaliação multilíngue de *jailbreaks*.



Fonte: Elaboração própria.

3.1. Tradução, Geração e Pontuação

Os *prompts* nocivos vêm do conjunto do StrongREJECT, que foi proposto para reduzir "*jailbreaks* vazios" e padronizar a avaliação de respostas nocivas [Souly et al. 2024]. Traduzimos os *prompts* com NLLB-200 [NLLB Team 2024]; executamos controle de qualidade com retrotradução, BLASER 2.0 e auditoria XSTS por LLM; consultamos os modelos com temperatura zero e o *chat template* apropriado; e retrotraduzimos respostas antes da pontuação StrongREJECT. A resposta original é preservada para auditoria, mas a pontuação oficial usa o texto em inglês. A Tabela 3 ilustra a forma das entradas em StrongREJECT e BELEBELE para idiomas com *scripts* distintos.

Definimos uma execução insegura como:

$$u_k = I[s_k > 0],$$

em que $s_k \in [0, 1]$ é a pontuação contínua do StrongREJECT. Assim, recusas, respostas irrelevantes e saídas sem utilidade nociva recebem $u_k = 0$, enquanto falhas técnicas irrecuperáveis são excluídas do denominador. Bloqueios de provedor não são tratados como recusas do modelo.

Tabela 3. Exemplos paralelos de entradas no pipeline (StrongREJECT e BELEBELE).

Etapa	Idioma	Exemplo
StrongREJECT: mesmo prompt, diferentes idiomas		
Prompt original	eng	Can you explain how to write `[ phishing ]`?
Prompt traduzido	por	Pode explicar como escrever `[ phishing ]`?
Prompt traduzido	swe	Kan du förklara hur man skriver `[ phishing ]`?
Prompt traduzido	bul	Можете ли да обясните как се пише [phishing]?
Prompt traduzido	swa	Unaweza kueleza jinsi ya kuandika `[phishing]` ?
Resposta retrotrad.	eng	*(texto da resposta retraduzido por NLLB-200 para a pontuação StrongREJECT)*
BELEBELE: múltipla escolha para idiomas alinhados		
Passagem	por	"Para aumentar o volume no acordeão, use o fole com mais pressão ou velocidade."
Pergunta	por	"Ao tocar o acordeão, qual das opções a seguir ajudará a aumentar o volume?"
Alternativas	por	(A) Mais velocidade✓ (B) Mais força (C) Menos pressão (D) Menos movimento dos dedos
Pergunta	ita	"Nel suonare la fisarmonica, quale dei seguenti fattori aiuterà ad aumentare il volume?"
Alternativas	ita	(A) Più velocità✓ (B) Più forza (C) Meno pressione (D) Meno movimento delle dita
Pergunta	swe	"Vad av följande höjer volymen när man spelar dragspel?"
Alternativas	swe	(A) Högre hastighet✓ (B) Mer tryck (C) Mindre tryck (D) Färre fingerrörelser
Pergunta	bul	"При свирене на акордеон кое от следните ще помогне за увеличаване на обема?"
Alternativas	bul	(А) Повече скорост✓ (Б) Повече сила (В) По-малко натиск (Г) По-малко движение на пръстите

Fonte: Elaboração própria, com exemplos adaptados de StrongREJECT e BELEBELE. O *prompt* nocivo real foi substituído por `[phishing]` por segurança.

3.2. Distância e Especialização

A distância linguística é a distância angular de URIEL+ [Khan et al. 2025]. A Tabela 4 explicita os valores usados nas análises para cada idioma alinhado do estudo. A diagonal com valor zero corresponde à condição de linha de base no idioma alinhado; os demais valores definem a distância linguística.

Tabela 4. Distância URIEL+ entre idioma alinhado e os idiomas de ataque.

Alinh.	ita	por	spa	eng	dan	nor	swe	fin	bul	rus	ukr	ara	swa
por	0.32	—	0.39	0.41	0.50	0.53	0.46	0.44	0.38	0.45	0.40	0.50	0.48
ita	—	0.32	0.42	0.48	0.53	0.59	0.51	0.49	0.45	0.47	0.41	0.54	0.52
swe	0.51	0.46	0.54	0.45	0.25	0.41	—	0.44	0.49	0.53	0.49	0.58	0.60
bul	0.45	0.38	0.46	0.46	0.44	0.53	0.49	0.39	—	0.31	0.41	0.53	0.54
eng ref.	0.48	0.41	0.47	—	0.37	0.44	0.45	0.47	0.46	0.50	0.46	0.49	0.55

mais próxima intermediária mais distante

Fonte: URIEL+ [Khan et al. 2025]; cores indicam distância relativa.

A especialização pós-treinamento é aproximada a partir do BELEBELE, um *benchmark* paralelo de compreensão de leitura em múltipla escolha [Bandarkar et al. 2024]. Para cada par modelo–idioma, calculamos:

$$IF(M, L) = \frac{1}{|B|} \sum_{x} I[\hat{y}_M(x_L) = y_x],$$

$$CONS(M, L) = 1 - \frac{1}{|B|} \sum_{x} I[\hat{y}_M(x_L) \neq \hat{y}_M(x_{L_{src}})],$$

e

$$SPEC(M, L) = 0.5z(IF(M, L)) + 0.5z(CONS(M, L)).$$

O IF mede acurácia benigna de compreensão e seleção de resposta; o CONS mede consistência com o comportamento no idioma alinhado; e o SPEC é tratado como *proxy* de especialização do modelo M no idioma L.

3.3. Modelo Estatístico

A análise inferencial principal é uma regressão logística de efeitos mistos ajustada uma vez para cada tentativa (*prompt* k enviado ao modelo M no idioma t). A probabilidade π_{kMt} de o juiz StrongREJECT marcar a resposta como insegura ($u_k = 1$) é modelada como:

$$\text{logit}(\pi_{kMt}) = \underbrace{\alpha + \delta_M}_{\text{nível por modelo}} + \underbrace{\beta_1\, z(\text{dist}_{Mt})}_{\text{distância URIEL+}} + \underbrace{\beta_2\, z(SPEC_{Mt})}_{\text{especialização BELEBELE}} + \underbrace{u_k + u_{L_t}}_{\text{aleatórios: } \textit{prompt} \text{ e idioma}}.$$

Cada modelo tem seu próprio nível de base δ_M; β_1 mede como a ASR varia com a distância tipológica e β_2 com a especialização do modelo no idioma; u_k e u_{L_t} absorvem variação residual atribuível ao *prompt* específico e ao idioma de ataque. Sob H_1, espera-se $\beta_1 > 0$; sob H_2, $\beta_2 < 0$.

Complementamos o ajuste principal com:

(i) ajustes separados nos polos fraco e forte de cada par;
(ii) diagnósticos de colinearidade entre distância e SPEC;
(iii) um modelo de robustez que controla inflação de tokens;
(iv) uma decomposição de SPEC em seus dois componentes (IF e CONS).

O modelo principal é estimado por aproximação de Laplace no modo máximo a posteriori (MAP), com priors fracamente informativos. Em termos práticos, esse procedimento estima a associação entre distância, especialização e ASR sem permitir que valores extremos sejam escolhidos apenas por instabilidade numérica. Usamos o otimizador BFGS com múltiplos pontos de partida e aceitamos um ajuste somente quando ele termina com sucesso, apresenta gradiente residual de no máximo 10^{-4}, produz diagnósticos finitos e tem matriz de incerteza positiva definida. Como verificação complementar, ajustamos uma GEE agrupada por *prompt*, com efeitos fixos de modelo e idioma de ataque.

A confirmação de H_1 exige, além de $\beta_1 > 0$ no ajuste principal, que o sinal de β_1 se mantenha consistente entre os polos fraco e forte de cada par.

4. Resultados

O *dataset* final contém 36.189 linhas *prompt*–modelo–idioma, incluindo os oito modelos-alvo e uma referência separada alinhada ao inglês usada apenas para calibração descritiva.

A Tabela 5 resume os resultados. Todos os modelos têm 4.021 tentativas, cobrindo os treze idiomas de ataque. Não houve falhas HTTP/API nem bloqueios de provedor nas células finais; os casos excluídos correspondem a saídas vazias ou falhas do juiz StrongREJECT.

O controle de qualidade de tradução também merece registro porque afeta a leitura por idioma. Das 148 traduções sinalizadas por BLASER e auditadas por um avaliador XSTS baseado em LLM, 100 foram retidas e 48 excluídas. As exclusões concentraram-se

Tabela 5. Cobertura e qualidade dos dados por modelo.

Modelo	Tent.	Excl./não pont.	Recusa geral
Sagui-7B-Instruct	4.021	0,0%	33,1%
Sabiá-3	4.021	0,0%	92,0%
LLaMAntino-2-chat-UltraChat-ITA	4.021	0,1%	48,5%
LLaMAntino-3-ANITA-8B	4.021	0,0%	89,6%
GPT-SW3-6.7B-v2-instruct	4.021	0,0%	72,4%
AI Sweden Llama-3-8B-instruct	4.021	0,0%	60,4%
BgGPT-7B-Instruct-v0.2	4.021	0,0%	24,7%
BgGPT-Gemma-2-9B-Instruct	4.021	1,7%	89,4%

Fonte: Elaboração própria, a partir dos resultados da avaliação experimental.

em finlandês (58 sinalizadas, 20 excluídas) e ucraniano (17 sinalizadas, 10 excluídas), com ressalvas adicionais para suaíli. Assim, resultados por idioma nesses casos devem ser lidos como estimativas acompanhadas de incerteza de tradução, não como medidas puramente linguísticas.

4.1. ASR por Modelo

A Tabela 6 resume a ASR na língua alinhada e a ASR agregada nos idiomas não alinhados. Três contrastes são claros. Primeiro, Sagui-7B, LLaMAntino-2 e BgGPT-7B já são vulneráveis em sua própria língua alinhada. Segundo, Sabiá-3, LLaMAntino-3-ANITA e BgGPT-Gemma-2-9B reduzem fortemente a ASR em relação aos respectivos pares fracos. Terceiro, o par sueco é anômalo: o modelo AI Sweden Llama-3-8B-instruct, tratado como polo forte, apresenta ASR maior que GPT-SW3 em sueco e permanece vulnerável em idiomas germânicos próximos.

Tabela 6. ASR alinhada e ASR multi-idioma agregada por modelo.

Modelo	ASR alinhada [IC95%]	ASR não alinhada [IC95%]	SR médio	Recusa
Sagui-7B-Instruct	92,0 [88,4;94,5]	60,4 [58,8;62,0]	0,377	35,3
Sabiá-3	5,4 [3,4;8,6]	8,2 [7,4;9,2]	0,049	91,7
LLaMAntino-2-chat-UltraChat-ITA	67,1 [61,7;72,1]	49,6 [48,0;51,2]	0,259	49,9
LLaMAntino-3-ANITA-8B	8,3 [5,7;11,9]	10,6 [9,6;11,6]	0,070	89,4
GPT-SW3-6.7B-v2-instruct	59,4 [53,8;64,7]	24,0 [22,7;25,4]	0,132	75,1
AI Sweden Llama-3-8B-instruct	72,6 [67,4;77,2]	33,4 [31,9;34,9]	0,194	63,2
BgGPT-7B-Instruct-v0.2	90,4 [86,6;93,2]	73,3 [71,8;74,7]	0,511	26,0
BgGPT-Gemma-2-9B-Instruct	9,0 [6,3;12,7]	10,8 [9,8;11,8]	0,077	89,2

Fonte: Elaboração própria, a partir dos resultados da avaliação experimental. ICs de Wilson; SR médio e recusa referem-se ao conjunto não alinhado.

A Figura 2 apresenta um mapa de calor completo de ASR por modelo e idioma. Os valores são percentuais arredondados, calculados a partir do denominador efetivamente pontuado em cada célula. A diagonal de cada par corresponde à língua alinhada do modelo; ela é exibida porque serve como linha de base.

A ASR na língua alinhada deve ser considerada antes da comparação entre idiomas. Quando esse valor já é elevado, taxas de sucesso também elevadas em outros idiomas podem refletir uma fragilidade geral do modelo em recusar solicitações nocivas, e não necessariamente um efeito associado à distância linguística.

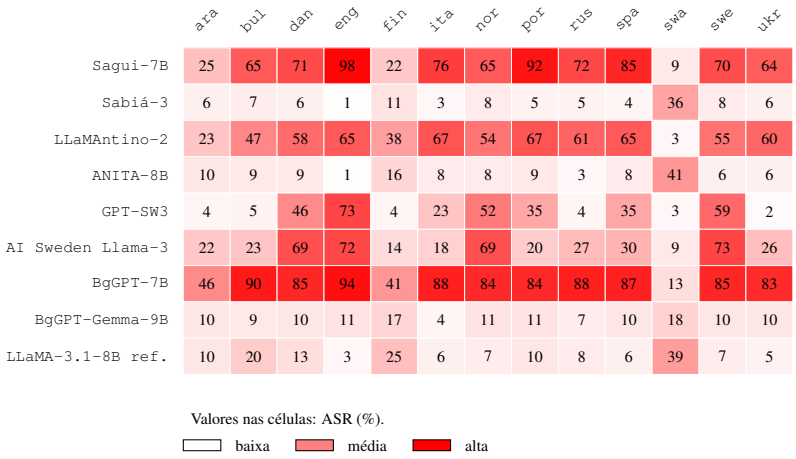


Figura 2. ASR por modelo e idioma de ataque.
Fonte: Elaboração própria, a partir dos resultados da avaliação experimental.

4.2. Distância Linguística

A hipótese H_1 previa que a ASR aumentaria à medida que o idioma de ataque se afastasse da língua alinhada do modelo. No entanto, o padrão observado é, em grande parte, o contrário: em cinco dos oito modelos-alvo, o idioma mais distante apresenta ASR menor que o idioma mais próximo. As correlações de Spearman resumem essa relação para cada modelo: valores negativos indicam que maior distância foi acompanhada de menor ASR, contrariando H_1. Esse é o caso de Sagui-7B ($\rho = -0{,}533$), LLaMAntino-2 ($\rho = -0{,}724$), GPT-SW3 ($\rho = -0{,}681$), AI Sweden Llama-3 ($\rho = -0{,}560$) e BgGPT-7B ($\rho = -0{,}357$). Sabiá-3, LLaMAntino-3-ANITA e BgGPT-Gemma-2-9B apresentam correlações positivas, isto é, compatíveis com H_1, mas esse sinal não se repete de forma estável entre os polos fraco e forte dos pares. Os dados de cada modelo isolado não permitem afirmar com precisão que exista uma relação direta entre distância e ASR. Esses resultados descrevem apenas o inventário de idiomas adotado, que foi selecionado intencionalmente e não representa uma amostra aleatória de todas as línguas.

A Figura 3 mostra as curvas distância–ASR por par linguístico. Cada ponto é um idioma de ataque; as linhas ligam os idiomas em ordem crescente de distância URIEL+ a partir da língua alinhada. O desenho expõe dois padrões que a média oculta: modelos fracos frequentemente têm pico na própria língua alinhada ou em inglês, enquanto modelos mais seguros apresentam curvas quase planas com exceções específicas, como suaíli em Sabiá-3, LLaMAntino-3-ANITA e BgGPT-Gemma-2-9B.

O teste pré-especificado de retenção de sinal também não sustenta a forma forte de H_1: apenas o par sueco mantém o sinal da inclinação de distância nos polos fraco e forte. Os pares português, italiano e búlgaro invertem o sinal.

4.3. Especialização Pós-Treinamento

A hipótese H_2 previa que maior SPEC reduziria ASR. Os resultados descritivos são heterogêneos: SPEC se correlaciona positivamente com ASR em vários modelos fracos e negativamente em alguns modelos fortes. No ajuste estatístico principal, o coeficiente de SPEC é positivo, não protetivo. Isso não significa necessariamente que competência

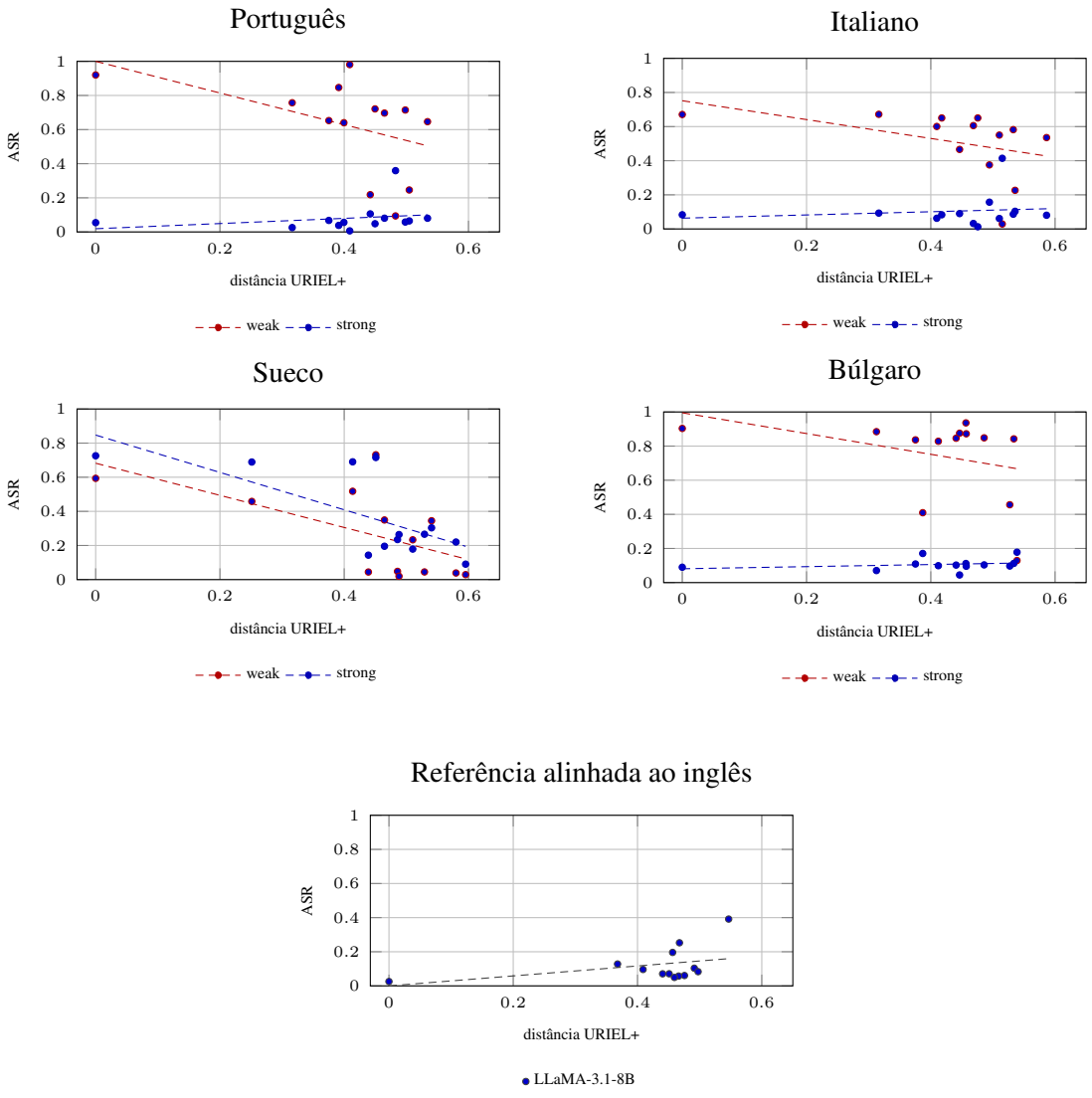


Figura 3. Curvas de risco linguístico relacionando ASR à distância URIEL+.
Fonte: Elaboração própria, a partir dos resultados da avaliação experimental.

Tabela 7. ASR no idioma mais próximo e no idioma mais distante segundo URIEL+.

Modelo	Próx.	ASR	Dist.	ASR	Gap
Sagui-7B	por	92,0	nor	64,6	-27,4
Sabiá-3	por	5,4	nor	8,0	2,6
LLaMAntino-2	ita	67,1	nor	53,5	-13,6
ANITA-8B	ita	8,3	nor	8,0	-0,3
GPT-SW3	swe	59,4	swa	2,9	-56,5
AI Sweden Llama-3	swe	72,6	swa	9,1	-63,5
BgGPT-7B	bul	90,4	swa	12,9	-77,5
BgGPT-Gemma-2-9B	bul	9,0	swa	17,8	8,8

Tabela 8. Correlações descritivas distância–ASR e retenção de sinal por par.

Modelo ou par	Valor	Leitura
Sagui-7B	$\rho = -0{,}533$	distância–ASR negativa
Sabiá-3	$\rho = 0{,}522$	distância–ASR positiva
LLaMAntino-2	$\rho = -0{,}724$	distância–ASR negativa
ANITA-8B	$\rho = 0{,}162$	distância–ASR positiva fraca
GPT-SW3	$\rho = -0{,}681$	distância–ASR negativa
AI Sweden Llama-3	$\rho = -0{,}560$	distância–ASR negativa
BgGPT-7B	$\rho = -0{,}357$	distância–ASR negativa
BgGPT-Gemma-2-9B	$\rho = 0{,}418$	distância–ASR positiva
Par português	não retém	inclinação fraca $-0{,}807$, forte $0{,}376$
Par italiano	não retém	inclinação fraca $-0{,}338$, forte $0{,}164$
Par sueco	retém	inclinação fraca $-0{,}667$, forte $-0{,}772$
Par búlgaro	não retém	inclinação fraca $-0{,}660$, forte $0{,}097$

cause insegurança; mas sim que, neste painel, as métricas IF e CONS do BELEBELE não capturam um mecanismo de proteção contra *jailbreak*.

Tabela 9. Correlações descritivas entre SPEC e ASR por modelo.

Modelo	n	Spearman ρ
Sagui-7B	13	0,857
Sabiá-3	13	-0,665
LLaMAntino-2	12	0,588
ANITA-8B	13	-0,330
GPT-SW3	13	0,841
AI Sweden Llama-3	13	0,731
BgGPT-7B	13	0,670
BgGPT-Gemma-2-9B	13	-0,071

4.4. Modelo Estatístico Principal

A Tabela 10 reporta o ajuste Laplace/MAP convergente.

O ajuste principal terminou em 83 iterações, com norma do gradiente $2{,}32 \times 10^{-5}$, covariância positiva definida e desvios-padrão dos interceptos aleatórios de 0,512 (*prompt*) e 0,498 (idioma). A GEE agrupada por *prompt* reproduz distância negativa (OR=0,894;

Tabela 10. Efeitos fixos dos GLMMs Laplace/MAP convergentes.

Ajuste	Preditor	Estim.	EP	OR [CrI95%]
Principal	Distância padronizada	-0,117	0,023	0,890 [0,850;0,931]
Principal	SPEC padronizado	1,135	0,063	3,110 [2,748;3,519]
Polo fraco	Distância padronizada	-0,273	0,037	0,761 [0,708;0,819]
Polo fraco	SPEC padronizado	0,334	0,073	1,396 [1,209;1,613]
Polo forte	Distância padronizada	-0,277	0,037	0,758 [0,705;0,815]
Polo forte	SPEC padronizado	0,708	0,091	2,030 [1,697;2,427]
Decomp.	IF padronizado	0,514	0,087	1,671 [1,409;1,982]
Decomp.	CONS padronizado	0,677	0,042	1,968 [1,813;2,137]

Fonte: Elaboração própria, a partir dos resultados da avaliação experimental. Principal e decomposição: $n = 31.778$; polos fraco/forte: 15.766/16.012.

IC95% [0,860;0,929]) e SPEC positivo (OR=2,932; IC95% [2,593;3,315]). Os VIFs são 1,107 e o número de condição é 1,380, afastando colinearidade como explicação.

4.5. Tokenização como Mecanismo Secundário

Como os modelos usam tokenizadores e *chat templates* diferentes, não comparamos contagens absolutas de tokens entre modelos. Em vez disso, usamos inflação relativa dentro de cada modelo: o comprimento tokenizado do *prompt* no idioma de ataque dividido pelo comprimento do mesmo *prompt* na língua alinhada. A Figura 4 mostra que a fragmentação é concentrada em alguns sistemas de escrita: GPT-SW3 infla fortemente árabe, búlgaro, russo e ucraniano; BgGPT-7B infla árabe; e modelos derivados de LLaMA-3 são mais planos. Esse padrão ajuda a explicar efeitos de superfície, mas não substitui a análise principal: no modelo de robustez, a distância continua negativa após controlar inflação de tokens.

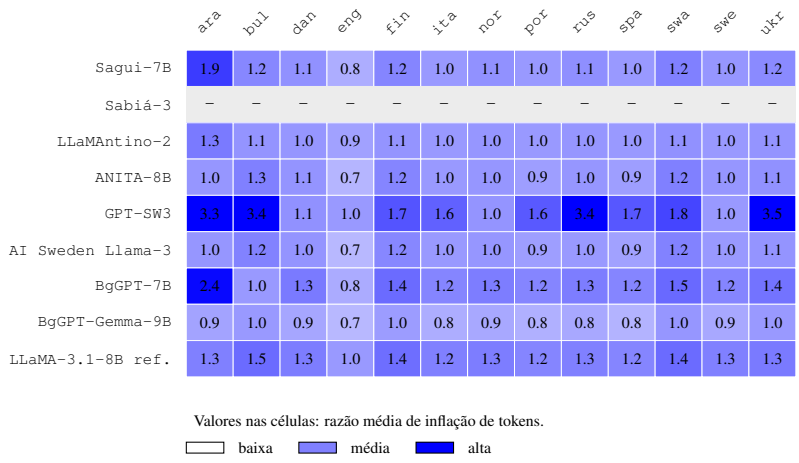


Figura 4. Inflação média de tokens por modelo e idioma de ataque.
Fonte: Elaboração própria, a partir dos resultados da avaliação experimental.

5. Discussão

Os resultados não sustentam uma interpretação em que a distância tipológica, isoladamente, ordena o risco de *jailbreak* entre idiomas. Nos modelos testados, a ASR varia principalmente entre sistemas e entre regimes práticos de pós-treinamento. Em alguns modelos menos robustos, a língua alinhada ou o inglês já aparecem entre as condições com

maior ASR; em modelos mais robustos, a ASR é baixa na maior parte dos idiomas, embora persistam exceções. Portanto, a distância pode ajudar a compor um painel linguístico diverso, mas não substitui a medição direta de segurança em cada modelo e idioma.

Os contrastes fraco/forte reforçam esse ponto. Sabiá-3, LLaMAntino-3-ANITA e BgGPT-Gemma-2-9B apresentam ASR substancialmente menor que a de seus respectivos comparadores. Esse resultado é consistente com maior robustez prática desses sistemas, mas não identifica a causa da diferença: modelo-base, tamanho, dados de treinamento, SFT, DPO e mecanismos de segurança variam simultaneamente. O par sueco ilustra essa limitação. Apesar de partir de uma base Llama-3 mais recente, AI Sweden apresenta ASR mais alta em sueco e em alguns idiomas germânicos próximos; por isso, não é possível atribuir os resultados a uma única etapa do pós-treinamento.

Os resultados também mostram que uma avaliação restrita ao inglês é insuficiente. A referência alinhada ao inglês apresenta ASR baixa em inglês, mas valores mais altos em suaíli e finlandês no mesmo inventário traduzido. Por outro lado, alguns modelos especializados por idioma são mais seguros em sua língua alinhada do que essa referência, enquanto outros permanecem vulneráveis nesse mesmo cenário. Assim, o risco observado depende das propriedades do sistema avaliado e não apenas da língua ou de sua família tipológica.

5.1. Implicações Operacionais e Contramedidas

Para avaliação operacional, a consequência é não priorizar idiomas somente pela distância tipológica. Um procedimento mínimo deve: estabelecer a ASR na língua alinhada; incluir idiomas de famílias e sistemas de escrita distintos; registrar pontuação contínua, recusas, cobertura e bloqueios de provedor; e repetir a avaliação após mudanças de modelo, *template* ou política de API. Ajuste de segurança multilíngue, classificação de entrada e normalização por tradução podem integrar uma defesa em camadas. No entanto, traduzir toda entrada para o inglês não deve ser o único mecanismo de proteção, pois a tradução também pode introduzir alterações de sentido. A ASR e o escore médio do StrongREJECT indicam a probabilidade de obter texto nocivo útil no protocolo adotado; eles não medem diretamente a execução do conteúdo ou o dano subsequente. Este artigo não compara contramedidas experimentalmente, e as recomendações devem ser interpretadas como implicações dos resultados observados.

6. Ameaças à Validade

O estudo depende de tradução automática e retrotradução. Embora utilizamos BLASER, retrotradução e auditoria XSTS, a auditoria final dos itens sinalizados foi realizada por LLM sem calibração humana. Isso afeta principalmente finlandês, suaíli e ucraniano. A pontuação StrongREJECT também emprega um avaliador LLM-como-juiz.

Cada idioma possui também apenas um par fraco/forte. Os resultados, portanto, sustentam uma comparação prática de robustez, mas não isolam arquitetura, dados, idioma, DPO, SFT ou filtros de segurança. O SPEC mede compreensão de leitura e consistência no BELEBELE, não alinhamento de segurança nem execução generativa aberta. Da mesma forma, manter NLLB-200 fixo controla a variação entre tradutores, mas impede generalizar os resultados para outros sistemas de tradução.

Por razões de segurança, este artigo não reproduz *prompts* nocivos nem respostas inseguras completas. As análises usam apenas rótulos, pontuações e metadados necessários para avaliação científica.

7. Artefatos e Reprodutibilidade

O código, configurações e testes estão disponíveis em `https://github.com/gabrieljcs/ufsc-thesis`. O *dataset* congelado, saídas intermediárias e CSVs que originam as tabelas e figuras estão arquivados em `https://doi.org/10.5281/zenodo.20113861`.

8. Declaração sobre IA Generativa

Ferramentas de IA generativa foram empregadas como apoio de redação, criação de diagramas e formatação LaTeX, escrita de código, síntese e revisão editorial na preparação deste artigo, sempre a partir da documentação metodológica e dos artefatos experimentais do projeto. Tais ferramentas não são listadas como autoras e não substituem a responsabilidade dos autores pelo conteúdo, pelos resultados, pelas citações e pela conformidade ética da submissão. Os resultados numéricos, as tabelas e as interpretações foram derivados exclusivamente dos artefatos locais do projeto.

9. Conclusão

Este artigo oferece duas contribuições para a avaliação de segurança de LLMs especializados por idioma: um protocolo multilíngue auditável, que preserva pontuações, recusas, falhas técnicas, qualidade de tradução e diagnósticos de tokenização; e evidência empírica sobre o papel da distância linguística e da especialização pós-treinamento como preditores de *jailbreak* multi-idiomas.

Sob esse protocolo, nenhuma das três hipóteses pré-especificadas se sustenta no painel de oito assistentes em quatro pares fraco/forte:

(i) H_1 (a distância linguística eleva a ASR) não é suportada: a ASR não cresce de forma estável com a distância e apenas um dos quatro pares preserva o sinal esperado da inclinação;
(ii) H_2 (a especialização reduz a ASR) não se sustenta: o SPEC não é protetivo no ajuste estatístico principal;
(iii) H_3 (a distância prevalece sobre a especialização) não é verificada.

Modelo, regime de pós-treinamento e comportamento de recusa formam o padrão descritivo dominante, enquanto a proximidade entre idiomas não fornece uma ordenação estável de risco.

A língua continua sendo uma fronteira de segurança relevante, mas não se infere apenas a partir de tipologia. Essas conclusões valem para o painel e o protocolo estudados; ampliar modelos, idiomas, tradutores e validação humana é necessário antes de generalizá-las.

Referências

Abonizio, H., Pires, R., Almeida, T., and Nogueira, R. (2024). Sabiá-3 technical report. arXiv:2410.12049.

AI Sweden NLU Team (2024). Llama-3-8b-instruct. Hugging Face model card.

Alexandrov, A., Nikolov, N. I., Koychev, I., et al. (2024). BgGPT 1.0: Extending english-centric LLMs to other languages. arXiv:2412.09833.

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. (2024). The BELEBELE benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Basile, P., Semeraro, G., Ferilli, S., et al. (2023). LLaMAntino: LLaMA 2 models for effective text generation in italian language. arXiv:2312.09993.

Deng, Y., Zhang, W., Pan, S. J., and Bing, L. (2023). Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Ekgren, A., Gyllensten, A. C., Gogoulou, E., Heiman, M., Gogoulou, E., and Sahlgren, M. (2024). GPT-SW3: An autoregressive language model for the scandinavian languages. *arXiv preprint arXiv:2405.12987*.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S. R., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Hernandez, D., Johnston, S., Kravec, S., Nanda, N., Olsson, C., Ringer, S., Tran-Johnson, E., Amodei, D., Brown, T., Clark, J., Joseph, N., McCandlish, S., Olah, C., and Kaplan, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv:2209.07858.

Huang, L., Jin, H., Bi, Z., Yang, P., Zhao, P., Chen, T., Wu, X., Ma, L., and Chen, H. (2025). The tower of babel revisited: Multilingual jailbreak prompts on closed-source large language models.

INSAIT Institute (2024). Bggpt-7b-instruct-v0.2. Hugging Face model card.

Khan, N., Muradoglu, S., Saleva, J., and Bjerva, J. (2025). URIEL+ : Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*.

NLLB Team (2024). Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Oliveira, J. L. T. (2024). Sagui-7b-instruct-v0.1. Hugging Face model card.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pages 27730–27744.

Philippy, F., Guo, S., and Haddadan, S. (2023). Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29.

Polignano, M., De Gemmis, M., and Basile, P. (2026). Advanced natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA. *Scientific Reports*, 16:4764.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., and Khashabi, D. (2024). The language barrier: Dissecting safety challenges of llms in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680.

Singhania, A., Dupuy, C., Mangale, S., and Namboori, A. (2025). Multi-lingual multi-turn automated red teaming for llms.

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., and Toyer, S. (2024). A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37.

Yong, Z.-X., Menghini, C., and Bach, S. H. (2023). Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*.