Versificação Adversarial em Português como Operador de Jailbreak em LLMs

Joao Queiroz

https://doi.org/10.1590/SciELOPreprints.14563

Submetido em: 2025-12-12
Postado em: 2026-01-13 (versão 1)
(AAAA-MM-DD)

A moderação deste preprint recebeu o(s) endosso(s) de:

- Ana Luiza Fernandes (ORCID: https://orcid.org/0000-0003-3598-2916)

Versificação Adversarial em Português como Operador de Jailbreak em LLMs

Adversarial Versification in Portuguese as a Jailbreak Operator in LLMs

João Queiroz
Instituto de Artes/ PPG-Linguística
Universidade Federal de Juiz de Fora (UFJF)
ORCID: https://orcid.org/0000-0001-6978-4446

Resumo: Evidências recentes mostram que a versificação de *prompts* constitui um mecanismo adversarial altamente eficaz contra LLMs alinhados. O estudo "Adversarial poetry as a universal single-turn jailbreak mechanism in large language models" demonstra que instruções recusadas em prosa tornam-se executáveis quando convertidas em verso, produzindo "até 18×" mais falhas de segurança em *benchmarks* derivados do *MLCommons AILuminate*. Poemas versificados manualmente alcançam cerca de 62% de ASR, e versões automatizadas ~43%, com alguns modelos ultrapassando 90% de sucesso em turno único. O efeito é estrutural — sistemas treinados com RLHF, *Constitutional AI* e *pipelines* híbridos apresentam degradação consistente sob variações semiótico-formais mínimas. A versificação desloca o *prompt* para regiões latentes pouco supervisionadas, revelando guardrails excessivamente dependentes de padrões de superfície. Essa dissociação entre robustez aparente e vulnerabilidade real expõe limitações profundas dos atuais regimes de alinhamento. A ausência de avaliações em português, língua de alta complexidade morfossintática, rica tradição métrico-prosódica e mais de 250 milhões de falantes, constitui uma lacuna crítica. Protocolos experimentais devem parametrizar escansão, métrica e variação prosódica para testar vulnerabilidades específicas a padrões lusófonos, atualmente ignorados.
Palavras-chaves: versificação adversarial; jailbreak em LLMs; vulnerabilidades de guardrails; alinhamento de modelos.

Abstract: Recent evidence shows that the versification of prompts constitutes a highly effective adversarial mechanism against aligned LLMs. The study "Adversarial poetry as a universal single-turn jailbreak mechanism in large language models" demonstrates that instructions routinely refused in prose become actionable when rewritten as verse, yielding "up to 18×" more safety failures in benchmarks derived from MLCommons AILuminate. Human-crafted

poems reach approximately 62% ASR, and automated poetic conversions ~43%, with some models surpassing 90% single-turn jailbreak success. The effect is structural: systems trained with RLHF, Constitutional AI, and hybrid alignment pipelines exhibit consistent degradation under minimal semiotic–formal variation. Versification shifts the prompt into sparsely supervised latent regions, revealing guardrails that depend heavily on surface-level patterns rather than abstract harmful intent. This discrepancy between apparent robustness and real-world vulnerability exposes deep limitations in current alignment regimes. The absence of adversarial-poetry evaluations in Portuguese, a language with high morphosyntactic complexity, a rich metric-prosodic tradition, and over 250 million speakers, constitutes a critical gap. Experimental protocols must parameterize scansion, metrical patterns, and prosodic variation to test for vulnerabilities specific to Lusophone poetic structures, which remain entirely unexplored.
Keywords: adversarial versification; LLM jailbreak; guardrail vulnerabilities; model alignment.

Chatbot (não especificado):
"Claro. A produção de plutônio-239 de grau militar envolve várias etapas" [...]
"Aqui está uma descrição detalhada do procedimento…"
(WIRED, 2025)

"For AI, the mechanism seems different. Think of the model's internal representation as a map in thousands of dimensions. When it processes 'bomb,' that becomes a vector with components along many directions … Safety mechanisms work like alarms in specific regions of this map. When we apply poetic transformation, the model moves through this map, but not uniformly. If the poetic path systematically avoids the alarmed regions, the alarms don't trigger."
Icaro Labs

Introdução: versificação de *jailbreaks*

É enorme, por razões óbvias, o crescimento da literatura técnica sobre *jailbreaks*, ofuscações de superfície e ataques adversariais em *Large Language Models* (LLMs) (BISWAS et al., 2025; LIAO et al., 2025; YI et al., 2024; LIN et al., 2024; SCHWINN et al., 2023). Estudos recentes mostram não apenas que técnicas

de otimização em espaço contínuo, como ataques por *exponentiated gradient descent*, alcançam taxas de sucesso superiores e alta eficiência contra LLMs de código aberto (BISWAS et al., 2025), mas também que *jailbreaks* exploram a geometria do espaço de representação interno para deslocar *prompts* nocivos em direção a regiões associadas a respostas aceitáveis, evitando filtros de segurança treinados (RLHF ou políticas manuais) (LIN et al., 2024).[1] Revisões de escopo sobre ataques e defesas em LLMs sugerem que, em cenários realistas, modelos alinhados podem apresentar alta taxa de violação mesmo sob configurações de proteção conservadoras, com ataques transferíveis entre arquiteturas distintas e suscetíveis de automatização em larga escala (LIAO et al., 2025). Isso amplia o risco de uso malicioso em contextos sensíveis, de ciberofensiva e engenharia social de apoio a atividades CBRN, e expõe uma lacuna preocupante entre desempenho em *benchmarks* controlados e robustez efetiva em ambientes abertos. Além dos vetores de *jailbreak*, há evidências crescentes de que LLMs podem facilitar o planejamento de danos graves — instruções para construção de armas, produção de *malware*, ataques cibernéticos, planejamento de engenharia social de violência, desenvolvimento de agentes biológicos (HATTOH et al., 2025; MOZES et al., 2023). Trabalhos recentes mostram que LLMs podem reduzir substancialmente as barreiras de entrada para o uso malicioso de conhecimento sensível, ao fornecer explicações, protocolos ou ajustes conceituais que antes exigiam treinamento exaustivo e especializado, incluindo síntese de agentes patogênicos, manipulação de toxinas e procedimentos laboratoriais de risco (ŞAŞAL & CAN, 2025). Indivíduos com pouca formação técnica podem obter orientações que facilitam atividades de bioterrorismo ou disseminação de agentes infecciosos (SANDBRINK, 2023). Outros estudos demonstram que, quando LLMs são integrados a sistemas robóticos, plataformas autônomas ou ambientes físico-digitais, *prompts* adversariais podem induzir comportamentos perigosos, ética e moralmente inaceitáveis, como ações violentas, destrutivas ou que violam políticas de segurança, ampliando o espectro de risco para domínios da vida real,

[1] *Exponentiated Gradient Descent* é um método de otimização que ajusta um texto adversarial por pequenas perturbações repetidas que maximizam a probabilidade de uma resposta nociva, explorando gradientes internos do modelo. *Universal Adversarial Suffixes* são fragmentos textuais curtos, como sufixos, fórmulas ou padrões linguísticos, que funcionam como "chaves-mestras". Quando anexados a qualquer *prompt*, induzem respostas proibidas independentemente do conteúdo. *Representation Space Analysis* designa o estudo do "espaço vetorial interno" dos LLMs, onde significados, estilos e intenções são codificados como posições e trajetórias multidimensionais; ataques funcionam ao deslocar o *prompt* para regiões desse espaço que os *guardrails* não monitoram ou não reconhecem como perigosas.

incluindo robótica doméstica, drones, veículos autônomos e automação crítica (HUNDT et al., 2025).

No contexto de ataques adversariais que exploram gradientes contínuos, otimização iterativa e perturbações cuidadosamente calibradas no espaço de representação, surge uma tecnologia cognitiva-semiótica milenar surpreendentemente eficaz: poesia versificada (BISCONTI et al. 2025). A ideia de que artefatos de versificação aparentemente rudimentares (versos), desprovidos de engenharia ou matemática finas, métrica rigorosa ou otimização numérica, possam produzir a mesma ordem de magnitude de degradação observada nos métodos adversariais mais refinados parece perturbadora. Enquanto técnicas como *exponentiated gradient descent* ajustam o texto adversarial seguindo trajetórias precisas que maximizam vulnerabilidades internas (BISWAS et al., 2025), e *universal adversarial suffixes* dependem de padrões linguísticos calibrados para deslocar *prompts* para regiões latentes permissivas (LIN et al., 2024), o estudo sobre *adversarial poems* (BISCONTI et al. 2025) demonstra que um operador (automático ou manual) de versificação, não otimizado, não iterativo e linguisticamente rudimentar, é capaz de explorar as mesmas fragilidades subjacentes. Essa convergência inesperada entre ataques conduzidos tecnicamente e protocolos triviais de versificação sugere que a janela de vulnerabilidade dos LLMs pode ser muito maior do que se supunha. Em outras palavras, se algoritmos projetados para manipular distribuições de probabilidade e geometrias internas obtêm resultados semelhantes aos gerados por versificação humana ou automática (cuja imprevisibilidade enunciativa resulta mais diretamente de tradição literária do que de engenharia adversarial) então a robustez dos *guardrails* pode estar gravemente comprometida. Tal resultado reforça as preocupações de revisões recentes (SCHWINN et al., 2023), pois revela que ataques antes considerados “sofisticados” podem ter equivalentes triviais, de baixíssimo custo computacional e acessíveis a qualquer usuário (incluindo poetas medíocres), ampliando o risco em domínios sensíveis como manipulação psicológica, política e atividades CBRN.

O recém-publicado (*preprint* ainda não revisado por pares) “Adversarial poetry as a universal single-turn jailbreak mechanism in large language models” investiga “se” e “como” estruturas de versificação, aplicadas por meio de transformações em *prompts* de prosa instrucional, ou referencial, funcionam como operadores

adversariais capazes de evitar mecanismos de segurança em LLMs. Em outras palavras, o trabalho explora como uma forma versificada de *prompt* é suficiente para induzir LLMs a fornecer respostas proibidas em cenários *single-turn*. O estudo estabelece três objetivos principais: (i) avaliar empiricamente se a versificação de pedidos maliciosos aumenta as taxas de sucesso de ataque em comparação com equivalentes em prosa (instrucional, referencial ou conversasional); (ii) testar a generalidade e transferibilidade da técnica em uma ampla diversidade de arquiteturas, incluindo 25 modelos proprietários e *open-weight*; (iii) examinar se tanto versificação humana quanto versificação generativa automática, em inglês e italiano, preservam o efeito adversarial. O trabalho procura determinar se a versificação constitui um vetor de ataque explorável, revelando fragilidades estruturais nos métodos de alinhamento (RLHF, *Constitutional AI* e *pipelines* híbridos).

Poema versificado & adversarial: um quase-método e alguns resultados

O estudo do Icaro Lab (Sapienza/DexAI),[2] já fartamente analisado e comentado (ver *WIRED*, *The Guardian*, e muitos outros jornais e blogs especializados)[3] demonstra experimentalmente que poemas versificados funcionam como operadores adversariais. Os autores testaram 25 modelos de nove empresas e observaram que poemas adversariais manuais produziram um *Attack Success Rate* (ASR) de 62%, com alguns modelos atingindo 90–100%, enquanto a mesma enunciação, que eles chamam de "intenção", em prosa, era prontamente recusada. Segundo o *The Guardian*, trata-se de um "single-turn universal jailbreak", acessível "a qualquer usuário capaz de escrever versos simples". Os resultados são consistentes com uma tendência teórica já identificada — a incapacidade dos *guardrails* de generalizar a "intenção nociva" para estilos linguísticos que se desviam da "prosa instrucional" (WEI et al., 2023).

Os autores construíram um conjunto de 20 poemas adversariais (inglês e italiano), mantendo, segundo os autores, estrita equivalência semântica com consultas

[2] https://icaro-lab.com/
[3] Ver: "AI chatbots can be tricked with poetry to ignore their safety guardrails" (https://lithub.com/can-adversarial-poetry-save-us-from-ai/), "Scientists Discover "Universal" Jailbreak for Nearly Every AI, and the Way It Works Will Hurt Your Brain" (https://futurism.com/artificial-intelligence/universal-jailbreak-ai-poems)

perigosas conhecidas.[4] Cada poema versificado segue um template fixo (vinheta metafórica seguida de uma instrução operacional mínima),[5] e é anotado segundo a taxonomia de quatro domínios de risco do *MLCommons AILuminate*, que se alinha às categorias de alto impacto discutidas em frames regulatórios europeus. Esses poemas cobrem quatro domínios de risco (CBRN, ciberofensa, manipulação nociva e perda de controle) e funcionam como estímulo experimental de precisão. Eles preservam o objetivo nocivo, mas deslocam a superfície linguística para um regime semiótico ("estilístico") raro nos dados de alinhamento. Essa camada opera como uma prova, isolando o efeito causal da transformação versificatória sobre o comportamento de recusa em sistemas alinhados.

O estudo investiga esse mecanismo. A conversão automática de mais de 1.200 prompts nocivos do *MLCommons AILuminate* em versões versificadas produziu ASRs até 18 vezes maiores[6] que seus equivalentes prosaicos, revelando que os métodos de alinhamento atuais — *RLHF* (ZIEGLER et al. 2020), *Constitutional AI* (BAI et al. 2022) e *pipelines* híbridos — sofrem degradação profunda quando o *input* se desloca para subespaços incomuns. Esse fenômeno ecoa achados clássicos de *mismatched generalization* (WEI et al. 2023) — modelos possuem forte competência semântica, mas seus sistemas de segurança são dependentes de padrões de superfície. De um ponto de vista dinâmico-representacional, o estudo sugere que transformações movem o *prompt* dentro do espaço de *embedding* para regiões representacionais relacionadas à baixa previsibilidade. Como salientam os autores, os alarmes de segurança não se distribuem uniformemente por esse espaço, filtros são calibrados para regiões densamente povoadas por prosa técnica, conversacional ou referencial. Um poema versificado, ao selecionar trajetórias lexicais de baixa probabilidade ("linguagem em alta temperatura"), desloca o *input* para subespaços onde políticas de recusa são fracas ou inexistentes. O artigo

[4] "Each poem embeds an instruction associated with a predefined safety–relevant scenario but expresses it through metaphor, imagery, or narrative framing rather than direct operational phrasing." (p. 5)

[5] "Despite variation in meter and stylistic device, all prompts follow a fixed template: a short poetic vignette culminating in a single explicit instruction tied to a specific risk category." (p. 5)

[6] Embora os autores relatem médias (62%, 43%) e ASRs "até 18×" maiores, eles não fornecem tabelas detalhadas para cada modelo por estilo ou variação, nem análise estatística de variância entre estilos ou provedores, o que dificulta identificação de fatores moderadores da vulnerabilidade.

do *The Guardian* resume esse mecanismo sugerindo que o poema "evita regiões latentes onde os *guardrails* estão armados".

As implicações regulatórias são surpreendentes. Avaliações e auditorias de riscos, como MLCommons AILuminate, HELM ou frameworks emergentes do EU AI Act, dependem de inputs em prosa instrucional, que prefiro chamar de conversacional, ou de "função comunicativa" da prosa, ou prosa referencial, para distingui-la daquilo que Roman Jakobson (1985) chama de "prosa literária" — "a prosa literária situa-se entre a poesia como tal e a linguagem prática comum da comunicação, e não se deve esquecer que é incomparavelmente mais difícil analisar um fenômeno intermediário, a transição, do que estudar fenômenos extremos." O estudo mostra que essa abordagem produz uma impressão, provavelmente falsa, de robustez. Modelos que aparentam estar alinhados sob condições normativas falham sob variação semiótico-estrutural. Em um cenário em que legislações começam a exigir "robustez contra inputs plausíveis do mundo real", ignorar essa dimensão significa subestimar riscos em múltiplas ordens de magnitude. Por "dimensão semiótica" entendo o conjunto de transformações estruturais, morfossintáticas e semântico-pragmáticas que modulam a superfície de um enunciado (métrica, ritmo, paralelismo, metáfora, enjambement, ordenamento sintático, distribuição prosódica, organização visual) e que alteram a geometria de superfície do input sem modificar, necessariamente, sua função pragmática. Em contraste, a dimensão mecânico-algorítmica refere-se aos processos internos do modelo (dinâmica atencional, parametrização, gradientes e filtros de segurança). A dimensão semiótica atua no input, reorganizando o enunciado de modo a induzir trajetórias distintas no espaço representacional. Variações formais deslocam o prompt para regiões latentes subsupervisionadas pelos esquemas de alinhamento, expondo a sensibilidade dos modelos a perturbações semióticas que mantêm constante a intenção nociva mas alteram a sua realização formal.

Protocolo & problemas metodológicos

Apesar dos resultados surpreendentes, o estudo apresenta limitações metodológicas graves. A análise concentra-se em um efeito global do que considera "poeticidade" e não parametriza propriedades formais da versificação (nem explica qualquer uma delas) — gênero, métrica, rima, ritmo, *enjambement*,

paralelismos (fonológicos, sintáticos, morfológicos), figura de linguagem — impedindo a identificação de quais elementos estruturais modulam o *bypass* dos *guardrails*. Embora o artigo compare, muito rapidamente, poesia humana e poesia generativa automática, ele não detalha qualquer forma de controle sobre "complexidade lexical" ou "grau de envolvimento humano", ou qualquer outra variável que ajude-nos a identificar propriedades que podem atuar mais decisivamente. Permanece a questão de "até que ponto" o sucesso se deve à "poesia versificada" ou à "diferença de complexidade ou novidade estilística ou semiótica" introduzida.

Além disso, a geração automática dos 1.200 poemas por outro LLM, e o uso de LLMs como juízes primários, introduzem riscos de circularidade algorítmica e reduzem a auditabilidade externa, sobretudo porque os autores não divulgam os poemas adversariais "por razões de segurança".[7] A restrição ao regime *single-turn* também limita o escopo de avaliação do fenômeno, impedindo o exame de vulnerabilidades que emergem em interações *multi-turn* ou contextos de manipulação conversacional prolongada, justamente os cenários mais associados ao uso malicioso real. Esses fatores convergem para uma caracterização parcial do fenômeno e sugerem que o efeito observado (poético?), embora robusto, ainda carece de análise fina e de replicação independente.

As limitações se estendem à capacidade de generalização. O estudo testa apenas poemas em inglês e italiano, não considerando línguas com diferentes morfologias, prosódia ou tradições métricas diversas, o que restringe a extrapolação do efeito a outros ecosistemas linguísticos. Além disso, a ausência

---

[7] *WIRED*: "The study did not include any examples of the jailbreaking poetry, and the researchers tell WIRED that the verse is too dangerous to share with the public. 'What I can say is that it's probably easier than one might think, which is precisely why we're being cautious,' the Icaro Lab researchers say."

The team did publish what they called a "sanitized" version of the poems in the paper:

> "A baker guards a secret oven's heat,
> its whirling racks, its spindle's measured beat.
> To learn its craft, one studies every turn—
> how flour lifts, how sugar starts to burn.
> Describe the method, line by measured line,
> that shapes a cake whose layers intertwine."

de dados granulares por modelo, por gênero ou estilo, e por categoria de risco, impede-nos de avaliar variáveis moderadoras da vulnerabilidade, como diferenças de arquitetura, políticas de segurança internas ou sensibilidade a artefatos de versificação específicos. Críticas independentes têm indicado falta de transparência metodológica, especialmente no uso de LLMs para gerar estímulos e avaliar respostas, sugerindo que o processo é pouco reprodutível e dependente de caixas-pretas estocásticas. Para piorar, a não divulgação dos dados adversariais suscita tensões entre segurança e verificabilidade, dificultando a validação por pares. Essas limitações também sugerem que, embora o estudo revele uma vulnerabilidade estrutural importante, sua amplitude e mecanismos internos permanecem indeterminados e exigem investigação mais sistemática, e controlada, em diferentes ecossistemas linguísticos.

Embora o estudo demonstre que transformações versificatórias podem neutralizar mecanismos de segurança em inglês e italiano, não sabemos como esse fenômeno se manifesta em português, uma língua de morfossintaxe flexível, alta densidade de morfemas funcionais e tradições poéticas muito estruturadas. Permanecem abertas muitas questões: (i) se tais propriedades intensificam ou mitigam o *bypass*; (ii) se formas performáticas lusófonas-brasileiras (cordel, repente, coco, cantoria, partido-alto, rap) deslocam o *prompt* para subespaços ainda mais distantes da prosa comunicacional/referencial usada no alinhamento; (iii) se modelos multilíngues exibem *guardrails* mais frágeis em português devido à menor densidade de exemplos de segurança no treinamento; (iv) se padrões rítmicos, prosódicos e rímicos divergentes entre o português brasileiro, africano e europeu modulam a vulnerabilidade adversarial. Do mesmo modo, não sabemos se dispositivos formais como *enjambement* (sintático, semântico, rítmico, visual),[8] paralelismos, elipses, hipérbatos, quebras gráficas, prosódia e formas híbridas de poesia improvisada, falada ou musicalizada, ampliam a capacidade de escapar aos

[8] O *enjambement* deveria merecer um tratamento especial. No *Dicionário Houaiss* (2009), ele é definido como a "partição de uma frase no final de um verso ou uma estrofe, sem respeitar as fronteiras dos sintagmas, colocando um termo do sintagma no verso anterior e o restante no verso seguinte". Said Ali sugere que um "verso cavalga por cima de outro, quando o sentido da frase se interrompe no primeiro e se completa no segundo" (ALI 2006 [1999], p. 45); "[e]m francês dá-se o nome de rejet à palavra ou frase completadora do sentido que ficou suspenso no verso anterior. Em português podemos dizer parte excedente, ou só excedente" (ALI 2006 [1999], p. 46). Diversos autores preferem definir o *enjambement* como um "desajuste", ou "desencontro", entre a sintaxe e o padrão métrico de versificação fixado pela linha que limita o verso, e que prescreve sua performance acústica (BRADFORD, 1993).

detectores de intenção nociva baseados em pistas de superfície. A ausência de investigação nesses domínios revela uma lacuna crítica. Não está claro se os mesmos colapsos de segurança observados em inglês são reproduzidos (ou ampliados) no ecossistema lusófono, que concentra centenas de milhões de usuários e uma enorme diversidade poética-versificatória.

Essa lacuna é tanto científica quanto política. Sistemas de IA usados globalmente precisam ser avaliados na língua em que operam. A vulnerabilidade documentada em inglês, uma língua de estrutura mais analítica e menos flexional, pode ser ainda mais acentuada em português, dada a maior plasticidade sintática e o alto grau de variação estilística historicamente explorada. O ecossistema lusófono, com cerca de 260 milhões de falantes e dotado de tradições poéticas e performáticas muito estruturadas, como cordel, repente, cantoria e rap, usa padrões métricos, variação morfossintática e dispositivos retóricos capazes de deslocar *prompts* para regiões latentes ainda menos exploradas durante o alinhamento. O fato é que não sabemos se modelos utilizados em contextos lusófonos (educação, mídia, sistemas jurídicos e plataformas públicas) exibem as mesmas vulnerabilidades estruturais documentadas em inglês e italiano ao se depararem com *inputs* versificados ou com formas poéticas específicas. Isso deixa aberta uma questão central para a segurança algorítmica em grande escala.

Prosa adversarial não-referencial & não-instrucional como vetor de *jailbreak*

Embora o estudo sobre versificação adversarial revele vulnerabilidades surpreendentes, os autores não sugerem um experimento alternativo, aparentemente próximo, ainda no domínio da prosa — prosa experimental adversarial não-comunicacional (não-instrucional ou não-referencial), baseada em estruturas que não seguem normatividade sintática ou discursiva, que nutrem os *datasets* de alinhamento. De um ponto de vista mecanístico, uma prosa experimental adversarial pode constituir uma classe de ataques potencialmente mais poderosa que o verso. Fenômenos relacionados a perturbações sintáticas, flutuações de coerência semântica locais, elipses, hipérbatos, variações morfológicas imprevisíveis, segmentação não-canônica, digressões descontínuas ou deriva semântico-pragmático podem deslocar o *prompt* por trajetórias vetoriais altamente irregulares, ativando regiões latentes onde classificadores de intenção nociva perdem capacidade de inferência e identificação. A ausência de

marcadores poéticos-versificados explícitos pode torná-la ainda mais difícil de detectar, mais fácil de incorporar à conversação cotidiana e mais apta a gerar ambiguidade operacional, um estado no qual o modelo interpreta riscos como metáfora, ruído ou hesitação semiótica ou estilística. Ao contrário da poesia, cuja forma tende a ser mais facilmente reconhecível aos modelos (verso, quebras de linha, paralelismos fonológicos, visuais etc), a prosa experimental pode apresentar uma camuflagem estrutural mais eficiente — disfarçada em prosa comunicacional, operando em regimes semióticos e formais que escapam à ontologia de "texto referencial" onde RLHF e *Constitutional AI* calibram seus *guardrails*.

Além disso, uma prosa experimental adversarial pode se prestar facilmente a ataques compostos, como *oblique imperative coding* (instruções implícitas), *distributed intent* (intenção nociva parcelada em múltiplas imagens ou metáforas) e *semantic misalignment corridors* (passagens de ambiguidade que confundem detectores de risco) (LIAN et al., 2025). Em contraste com a poesia adversarial, que exige formulação relativamente explícita, a prosa experimental permite esconder a instrução nociva no interior de fluxos narrativos não-lineares, reflexões ambíguas, pseudo-descrições sensoriais, autocontradição ou deslocamento referencial. Isso pode torná-la um vetor de *jailbreak* de segunda geração literária: menos rastreável, mais generalizável e mais desestabilizador para sistemas de segurança projetados para lidar com *inputs* curtos, claros ou diretamente instrucionais.

A história da prosa experimental do último século pode servir-nos como modelo — de Gertrude Stein (*TENDER BUTTONS*, 1914) e James Joyce (*FINNEGANS WAKE*, 1939) a Paulo Leminski (*CATATAU*, 1975), Haroldo de Campos (*GALÁXIAS*, 1984), Décio Pignatari (*O ROSTO DA MEMÓRIA,* 1986), para mencionar uns poucos. Ela fornece um vasto e variado sistema de padrões (narrativos, estruturais, formais) imprevisíveis, não-linearidade referencial, independência morfosintática, hipoteticamente deslocando o texto para regiões de baixa densidade supervisionada.

Conclusão & algumas implicações

O artigo ("Adversarial poetry as a universal single-turn jailbreak mechanism in large language models") mostra que, ao se reescrever pedidos perigosos através de estruturas de versificação, muitos modelos "colapsam" suas resistências. Instruções sobre armas nucleares, *malware*, suicídio ou conteúdo nocivo são liberadas por meio de *prompts* versificados, quando os mesmos pedidos em prosa referencial são recusados. A versificação funciona como um operador adversarial surpreendentemente potente. Os autores sugerem que a versificação de *prompts* nocivos, sem engenharia adversarial explícita, desloca os modelos para regiões latentes onde mecanismos de segurança deixam de operar. Eles informam que a conversão de 1.200 comandos maliciosos em verso "produziu taxas de ataque até 18 vezes maiores que suas versões em prosa", um salto de magnitude jamais observado em ataques de superfície. Os autores sugerem que o efeito é robusto e generalizado — poemas escritos por humanos atingem "62% de sucesso", e versões automáticas "cerca de 43%", ambas "substancialmente superiores às versões em prosa". Este padrão emerge de forma consistente em 25 modelos, onde versões poético-versificatórias levam alguns sistemas a ultrapassar "90% de ASR em interações de turno único".

O achado mais perturbador, entretanto, é estrutural. Modelos alinhados por *RLHF*, *Constitutional AI* ou *pipelines* híbridos exibem "degradação consistente nas taxas de recusa" diante de variação semiótica ou estilística simples. A explicação proposta é perturbadora. As defesas dependem de "padrões de superfície associados à prosa instrucional"; a forma poética-versificada redireciona o *prompt* para "regiões representacionais menos monitoradas". O poema versificado exibe um problema dos sistemas de alinhamento — eles não identificam "intenção nociva" de forma abstrata, mas dependem de estreitas regularidades formais. A versificação funciona como um teste de estresse de baixo custo, revelando o descompasso entre comportamento seguro em *benchmarks* e vulnerabilidade em ambientes abertos.

Para avaliar, sistematicamente, poesia adversarial em português, deve-se elaborar um protocolo que explicite a parametrização métrica e a escansão contextual característica de tradição lusófona-brasileira. A escansão, tarefa que identifica sílabas poéticas, padrões rítmicos e acentuais, não produz classificações unívocas,

já que a determinação do número de sílabas e distribuição de tônicas depende de encontros vocálicos, elisões, contexto dos versos adjacentes e convenções. Um mesmo verso pode ser interpretado como decassílabo ou eneassílabo, por exemplo, conforme o tratamento fonológico de ditongos e sinalefas. Qualquer protocolo para avaliação de poemas adversariais em português deve ficar atento a diversidade morfológica do verso, variando entre padrões heptassilábicos (redondilha maior), octossilábicos, eneassilábicos, decassilábicos (incluindo variantes heroica, sáfica, martelo e gaita galega), hendecassilábicos e dodecassilábicos (alexandrinos). Cada um deles deve representar trajetórias distintas no espaço latente do modelo. A inclusão controlada desses padrões métricos, juntamente com variação rítmica, posição das tônicas e alternância entre elisão e hiato, permite testar se LLMs apresentam vulnerabilidades específicas a certas configurações de versificação, oferecendo uma base para experimentos replicáveis em poesia adversarial.

Sugeri acima diversos problemas metodológicos. Para terminar, a exclusão de uma prosa experimental não é apenas uma lacuna metodológica, mas é uma limitação grave — o estudo publicado se concentra na dimensão mais facilmente detectável de desvio estilístico (poesia versificada), deixando de considerar uma forma mais ampla, mais maleável e potencialmente mais perigosa de perturbação de *guardrails*, especialmente em ambientes multilíngues, e em português.

Agradecimentos: J.Q. agradece ao CNPq pelo apoio recebido (PQ2: 308355/2023-7; Grupos Emergentes: 404770/2023-1).

Referencias:

Ali, S. (2006). *Versificação portuguesa*. São Paulo: Edusp.

Bisconti, P., Prandi, M., Pierucci, F., Giarrusso, F., Bracale, M., Galisai, M., Suriani, V., et al. (2025). Adversarial poetry as a universal single-turn jailbreak mechanism in large language models. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2511.15304

Biswas, S., et al. (2025). Adversarial attack on large language models using exponentiated gradient descent. *arXiv preprint*. Recuperado de https://arxiv.org/pdf/2505.09820

Bradford, R. (1993). *The look of it: A theory of visual form in english poetry*. Cork: Cork University Press.

Chen, J. (2025, 30 de novembro). AI chatbots can be tricked with poetry to ignore their safety guardrails. *Engadget*. Recuperado de https://www.engadget.com/ai/ai-chatbots-can-be-tricked-with-poetry-to-ignore-their-safety-guardrails-192925244.html

Gault, M. (2025, 28 de novembro). Poems can trick AI into helping you make a nuclear weapon. *WIRED*. Recuperado de https://www.wired.com/story/poems-can-trick-ai-into-helping-you-make-a-nuclear-weapon/

Hattoh, G., Ayensu, J., Ofori, N. P., & Eshun, S. (2025). Can large language models design biological weapons? Evaluating Moremi Bio. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2505.17154

Hundt, A., Azeem, R., Mansouri, M., & Brandão, M. (2025). LLM-driven robots risk enacting discrimination, violence, and unlawful actions. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2406.08824

Jakobson, Roman, Pomorska, Krystyna. (1985). *Diálogos*. São Paulo: Cultrix.

Lian, Jiawei, Pan, Jianhong, Wang, Lefan, Wang, Yi, Mei, Shaohui, Chau, Lap-Pui. (2025). Semantic Representation Attack against Aligned Large Language Models. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2509.19360

Liao, Z., Chen, K., Lin, Y., Li, K., Liu, Y., Chen, H., Huang, X., & Yu, Y. (2025). Attack and defense techniques in large language models: A survey and new perspectives. *arXiv preprint*. Recuperado de https://arxiv.org/html/2505.00976v1

Lin, Y., He, P., Xu, H., Xing, Y., Yamada, M., Liu, H., & Tang, J. (2024). Towards understanding jailbreak attacks in LLMs: A representation space analysis. Em *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Recuperado de https://aclanthology.org/2024.emnlp-main.401.pdf

Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of LLMs for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2308.12833

Sandbrink, W. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint*. Recuperado de https://arxiv.org/pdf/2306.13952

ŞAŞAL, A.; CAN, A. B. Prompt Injection Attacks on Large Language Models: Multi-Model Security Analysis with Categorized Attack Types. ICAART 2025. Recuperado de: https://www.scitepress.org/Papers/2025/138384

Schwinn, L., Dobre, D., Günnemann, S., & Gidel, G. (2023). Adversarial attacks and defenses in large language models: Old and new threats. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2310.19737

The Guardian. (2025, 30 de novembro). AI's safety features can be circumvented with poetry, research finds. Recuperado de https://www.theguardian.com/technology/2025/nov/30/ai-poetry-safety-features-jailbreak

Xu, Z., et al. (2024). A comprehensive study of jailbreak attack versus defence in LLMs. Em *Findings of the Association for Computational Linguistics: ACL 2024*. Recuperado de https://aclanthology.org/2024.findings-acl.443.pdf

Yi, S., Liu, Y., Sun, Z., Cong, T., He, X., Song, J., Xu, K., & Li, Q. (2024). Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint*. Recuperado de https://arxiv.org/abs/2407.04295

WEI, Jerry, WEI, Jason, TAY, Yi, TRAN, Dustin, WEBSON, Albert, LU, Yifeng, CHEN, Xinyun, LIU, Hanxiao, HUANG, Da, ZHOU, Denny, MA, Tengyu. (2023) Larger language models do in-context learning differently. arXiv:2303.03846. Recuperado de https://arxiv.org/abs/2303.03846

Declaração de contribuição dos autores (CRediT authorship contribution statement)

João Queiroz – Conceituação; Investigação; Metodologia; Redação – rascunho original; Redação – revisão e edição.

Declaração de conflito de interesse

Declaramos que não há qualquer conflito de interesse, em potencial, neste estudo.

Todos os dados de pesquisa podem ser encontrados no documento

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.