Os LLMs compreendem expressões idiomáticas? Evidências para uma teoria da Competência Fraseológica Artificial Simulada

Thyago Jose da Cruz

https://doi.org/10.1590/SciELOPreprints.16937

Submetido em: 2026-07-16
Postado em: 2026-08-11 (versão 1)
(AAAA-MM-DD)

Os LLMs compreendem expressões idiomáticas? Evidências para uma teoria da Competência Fraseológica Artificial Simulada

Do LLMs Understand Idioms? Evidence for a Theory of Simulated Artificial Phraseological Competence

Thyago José da Cruz (UFMS/ PPGLetras-UEMS)

https://orcid.org/0000-0001-5562-8485

Resumo: Esta pesquisa examina, por meio de experimento quase controlado, a performance do modelo de linguagem de grande escala Claude Haiku 4.5 no processamento de expressões idiomáticas do português brasileiro. Fundamentado na articulação entre Fraseologia e Linguística Cognitiva, o estudo propõe o conceito de competência fraseológica artificial simulada, definida como o desempenho funcional baseado em regularidades estatístico-distribucionais, que não passaram pela experiência da corporalidade, vivência sociocultural e inferência pragmática como ocorre com os humanos. O corpus experimental, constituído por 140 unidades fraseológicas (expressões idiomáticas opacas, somáticas, culturais e contraprovas experimentais), foi submetido a três condições de *prompting* — *zero-shot, few-shot* contextualizado e *Chain-of-Thought* com *role-playing* — avaliando-se quatro dimensões: detecção de idiomaticidade, precisão semântica, adequação pragmática e sensibilidade cultural. Os dados obtidos demonstram elevada acurácia semântica das unidades convencionalizadas, mas desempenhos assimétricos nas dimensões pragmática e cultural, além de uma relevante alucinação figurativa diante de expressões idiomáticas inexistentes. A condição *Chain-of-Thought* demonstrou avanços qualitativos consistentes, reduzindo falsos positivos. Os resultados reforçam a hipótese de que o LLM analisado não dispõe de mecanismos equivalentes à competência fraseológica humana, embora haja evidências de que estratégias de engenharia de prompt possam mitigar parcialmente tais limitações arquiteturais.

Palavra-chave: fraseologia; competência fraseológica; modelos de linguagem em grande escala.

Abstract: This study examines, by means of a quasi-controlled experiment, the performance of the Claude 4.5 Haiku large language model in processing Brazilian Portuguese idiomatic expressions. Grounded in the intersection of Phraseology and Cognitive Linguistics, the study proposes the concept of simulated artificial phraseological competence, defined as a functional performance based on statistical-distributional regularities, devoid of the embodied experience, sociocultural immersion, and pragmatic inference characteristic of human language users. The experimental corpus, consisting of 140 phraseological units (including opaque, somatic, and cultural idioms, alongside experimental control samples), was evaluated across three prompting conditions—zero-shot, contextualized few-shot, and Chain-of-Thought with role-playing. Four dimensions were assessed: idiomaticity detection, semantic precision, pragmatic appropriateness, and cultural sensitivity. The findings demonstrate high semantic accuracy for conventionalized units, yet reveal asymmetrical performance across the pragmatic and cultural dimensions, alongside a pronounced tendency toward figurative hallucination when encountering non-existent idioms. The Chain-of-Thought condition yielded consistent qualitative improvements, reducing false positives. These results reinforce the hypothesis that the analyzed LLM lacks mechanisms equivalent to human

phraseological competence, although evidence suggests that prompt engineering strategies can partially mitigate these architectural limitations.

Keywords: phraseology; phraseological competence; large language models.

1 Introdução

Os Grandes Modelos de Linguagem (*Large Language Models*, LLMs) avançaram no Processamento de Linguagem Natural (PLN), podendo realizar tradução automática, geração de textos e raciocínios complexos. Contudo, modelos como ChatGPT, Claude, Gemini e Llama ainda enfrentam dificuldades com unidades fraseológicas, em especial com expressões idiomáticas (Cheng & Bhat, 2024)

As expressões idiomáticas, unidades da língua de caráter pluriverbal, conotativo e cristalizado (Xatara, 1998), possuem uma característica não composicional. Devido a essa natureza, os modelos de LLMs, cuja arquitetura se alicerça na tecnologia Transformer, tradicionalmente dependem de padrões distribucionais e estatísticos assimilados e aprendidos durante seu pré-treinamento. Para Rezaeimanesh *et al*. (2024), "apesar de sua prevalência na linguagem falada, os modelos de tradução automática (TA) de última geração têm dificuldades em traduzir expressões idiomáticas, muitas vezes interpretando-as literalmente como expressões composicionais de coocorrência" (Rezaeimanesh *et al*., 2024, p.1, tradução nossa).

Neste artigo, examinamos os principais desafios identificados na literatura científica atual no que se refere, principalmente, ao processamento de expressões idiomáticas por LLMs. Abordamos quatro pontos principais (falhas no uso de contexto para raciocínio metafórico; limitações na detecção de EI; problemas de alucinação metafórica e ancoragem cultural; e questões metodológicas relacionadas a *benchmarks* e métricas de avaliação), que são fundamentais para a argumentação e verificação de nossa pergunta de pesquisa: as LLMs possuem uma competência fraseológica operacional emergente, mas não uma competência fraseológica sociocognitiva plena como a dos humanos? Partimos da hipótese de que os humanos possuem uma competência fraseológica que está atrelada ao cognitivo, a experiências socioemocionais e culturais, enquanto as LLMs apresentam um desempenho funcional distribuído, ou seja, manipulam regularidades fraseológicas reais, porém não possuem conhecimento experiencial.

2 A natureza não composicional das Expressões Idiomáticas

As expressões idiomáticas distinguem-se por sua natureza não composicional: o significado do conjunto não resulta da simples soma de seus elementos constituintes. Um exemplo é a expressão "bater as botas", que significa "morrer" e cuja interpretação não guarda relação direta com as palavras que a compõem. Esse tipo de construção pode representar um verdadeiro calcanhar de Aquiles para as inteligências artificiais generativas.

Pesquisas recentes, como a de Li *et al*. (2024), evidenciam que as LLMs frequentemente falham em capturar o sentido não composicional das expressões idiomáticas (EIs), o que resulta

em interpretações literais equivocadas e traduções insatisfatórias. De modo complementar, (Cheng & Bhat, 2024) confirma essa dificuldade ao demonstrar que modelos generativos enfrentam obstáculos significativos ao lidar com EIs, justamente devido à sua natureza não composicional. Nesses casos, o significado torna-se de difícil apreensão quando não se considera a dimensão metafórica que lhes é inerente.

Os estudos de Li *et al*. e Cheng *et al*., se analisados em conjunto, demonstram uma padrão estrutural recorrente, isto é, quando solicitamos às LLMs que definam ou traduzam uma EI, elas podem falhar na não composicionalidade não pela ausência de dados sobre esses tipos de unidades fraseológicas em seu corpora de treinamento, mas devido ao seu processamento distribucional que opera fundamentalmente a partir de regularidades estatísticas de coocorrência lexical e não sobre a projeção metafórica que caracteriza o processamento cognitivo humano. Logo, essa dificuldade não é meramente quantitativa - se fosse, resolveríamos oferecendo às IAs generativas mais dados - mas são de natureza cognitiva e arquitetural. As LLMs carecem daquilo que a Linguística Cognitiva denomina de corporalidade (*embodiment*) e simulação mental (simulação corporificada).

Conforme Gibbs Jr & Macedo (2010), a corporalidade e a simulação mental constituem pilares fundamentais para a compreensão do funcionamento do pensamento e da linguagem. A corporalidade refere-se à natureza intrinsecamente vinculada ao corpo na cognição humana, manifestando-se em uma interação dinâmica na qual o pensamento emerge da relação entre cérebro, corpo e mundo. Essa base, derivada das experiências corporais, sustenta a elaboração de conceitos mais abstratos. Já a simulação mental corresponde ao mecanismo cognitivo que possibilita a compreensão da linguagem e dos conceitos por meio da capacidade de imaginar ações relevantes no mundo externo, do processamento de metáforas oriundas de experiências táteis, cinéticas e contínuas, e da categorização de objetos como representações das vivências associadas a eles — como no caso da palavra “cadeira”, que pode ser mobilizada para explicar a expressão idiomática “tomar um chá de cadeira”.

Tanto o processamento da corporalidade quanto o da simulação mental permitem que os seres humanos construam e atualizem a metaforicidade dos conceitos a partir de experiências sensório-motoras e socioculturais compartilhadas. Esse intervalo constitui o núcleo do que este trabalho denomina “competência fraseológica artificial simulada”, entendida como um desempenho funcional que imita os resultados da competência fraseológica humana, mas sem se apoiar nos processos sociocognitivos e socioculturais que a estruturam. No âmbito metodológico desta pesquisa, tal padrão reforça a necessidade de elaborar prompts que explicitem os domínios metafóricos envolvidos nas expressões idiomáticas, de modo a compensar, por meio da engenharia de prompt, a ausência de ancoragem experiencial nos LLMs.

3 Falhas no uso de Contexto: o paradoxo contextual

Uma das contribuições recentes da literatura científica é o chamado “paradoxo contextual”, que evidencia limitações no raciocínio metafórico realizado por LLMs. Cheng *et al*. (2024), por exemplo, ao conduzir diversos testes com modelos de linguagem pré-treinados de propósito geral baseados em *deep learning*, observou que esses sistemas apresentam desempenho inferior quando expostos a contextos circundantes em tarefas que envolvem idiomaticidade. Curiosamente, ao se retirar tais contextos, verificou-se uma melhora de até 3,89% no desempenho.

Complementarmente, Mi *et al.* (2025) acrescenta que LLMs como ChatGPT e Llama, em experimentos com conjuntos de dados contrastivos e controlados, demonstram falhas ao desambiguar leituras metafóricas e literais quando expostas a pistas contextuais, recorrendo mais a heurísticas superficiais e a estatísticas de verossimilhança das sentenças. Assim, para esses pesquisadores, o êxito dos modelos correlaciona-se diretamente com a probabilidade da sentença, em vez de resultar de um raciocínio contextual genuíno. Knietaite *et al*. (2024), por sua vez, observam que informações contextuais provenientes de sentenças circundantes podem, de forma paradoxal, distrair o modelo. Nesse sentido, equilibrar a qualidade dos dados e o acesso ao contexto torna-se essencial para a adequada compreensão da idiomaticidade por parte das LLMs.

Os estudos de Cheng *et al*. (2024), Mi *et al*. (2025) e Knietaite (2024) evidenciam que a presença de contextos — requisito fundamental para a ativação da competência fraseológica em humanos, em sua dimensão inferencial e pragmática — pode paradoxalmente atuar como fator de interferência e reduzir o desempenho das LLMs. Esses resultados indicam que o processamento contextual dos modelos tende a operar por meio de heurísticas estatísticas de verossimilhança, privilegiando a sentença mais provável em vez da mais adequada do ponto de vista idiomático. Em contraste, a competência fraseológica humana está vinculada às inferências sociocognitivas e socioculturais que os falantes mobilizam ao recorrer a conhecimentos enciclopédicos e a esquemas mentais de imagens para desambiguar leituras metafóricas das literais. Defendemos, portanto, que o processamento de simulação contextual sem uma interpretação pragmática efetiva pode inviabilizar o acerto das IAs generativas diante de unidades fraseológicas metafóricas. Tal limitação reforça a necessidade de elaborar prompts cuidadosamente estruturados em termos contextuais e semanticamente orientados, a fim de suprir, ainda que parcialmente, a ausência de inferência sociocognitiva genuína.

4 Limitações na detecção de Expressões Idiomáticas

A detecção de expressões idiomáticas permanece, por enquanto, uma tarefa desafiadora para as LLMs. Fornaciari *et al*. (2024) demonstram que esses modelos frequentemente apresentam dificuldade em distinguir significados literais de idiomáticos, atribuindo, em diversos casos, caráter idiomático a sequências de palavras aleatórias no texto. Isso se deve, segundo os autores, "os dados de pré-treinamento provavelmente contêm as PIEs com significado idiomático com mais frequência do que com significado literal, os modelos tendem a classificar essas expressões como idiomáticas em vez de literais com base na distribuição de probabilidade" (Fornaciari *et al*., 2024, p. 5, tradução do autor).

5 A alucinação figurativa e os problemas com culturas diferentes

Um dos desafios apontados pela literatura recente é a chamada "alucinação figurativa", que, segundo Hosseini *et al.* (2026), corresponde à tendência das LLMs de gerar ou validar expressões idiomáticas inexistentes em uma comunidade linguística, mas que parecem plenamente plausíveis. Esses pesquisadores argumentam que os modelos enfrentam dificuldades em distinguir

de forma confiável expressões idiomáticas autênticas e, com frequência, acabam alucinando durante a execução da tarefa, seja ao inventar novas EIs, seja ao fornecer explicações coerentes, porém falsas, acerca da realidade linguística humana.

A dificuldade das IAs generativas em compreender, adaptar-se e reproduzir diferentes contextos culturais manifesta-se de diversas formas. Um exemplo é apontado por Attia *et al*.(2026), que evidenciam as consideráveis limitações das LLMs ao lidar com expressões idiomáticas culturalmente marcadas. Além disso, a pesquisa demonstra um nível de eficiência que tende a favorecer culturas mais influentes, como a anglófona, em detrimento de outras tradições linguísticas.

Os estudos de Hosseini *et al*. (2026) e Attia *et al*. (2026) convergem ao demonstrar que as LLMs não apenas falham em reconhecer expressões idiomáticas autênticas, mas também tendem a produzir constantemente unidades fraseológicas inexistentes, validando-as com explicações coerentes, porém socioculturalmente falsas — fenômeno denominado "alucinação figurativa". Esse comportamento revela que os modelos carecem de um mecanismo de verificação vinculado às comunidades linguísticas reais, uma vez que não dispõem do conhecimento enciclopédico compartilhado nem do simbolismo culturalmente sedimentado Bertrán, (2008); Dobrovol'skij e Piirainen (2018). Tais dimensões funcionam, na competência fraseológica humana, como filtros que distinguem a expressão idiomática genuína da combinação livre meramente plausível e metafórica. Por isso, como será discutido nas seções seguintes, no plano metodológico, esse tipo de comportamento das LLMs justifica a necessidade de prompts experimentais com marcações culturais e contraexemplos interculturais, a fim de verificar em que medida a contextualização fornecida externamente pode atenuar a alucinação figurativa.

6 Os desafios das LLMs com línguas menos influentes

As limitações das inteligências artificiais tornam-se mais evidentes quando saímos do eixo das línguas mais influentes, isto é, advindas de países com mais poder aquisitivo e bélico. Estudos recentes, como o de Khoshtab *et al*. (2025), demonstram que os modelos de linguagem em grande escala apresentam dificuldades consideráveis ao analisar elementos comparativos (por exemplo, "como um touro") ou EIs ("boi de piranha"). Os modelos de código aberto performam ainda menos quando são comparações metafóricas ou expressões idiomáticas de línguas com menos dados na Internet, em comparação com as ditas "dominantes". Kim *et al*. (2025) corroboram com esses dados ao afirmar que os modelos geralmente memorizam mais expressões idiomáticas em inglês, alemão e chinês do que em coreano, turco ou árabe.

7 Novos testes: como medir o que as LLMs não entendem

Com o intuito de enfrentar as falhas que diferentes tipos de LLMs apresentam ao lidar com expressões idiomáticas, pesquisadores desenvolveram, entre 2022 e 2026, ferramentas de teste específicas, denominadas *benchmarks*. Entre elas, destacam-se: *rolling the Dice* (Mi *et al*., 2025), voltada para o uso de contextos na identificação e análise de EIs; *IdioTS*, programada para a

detecção de expressões idiomáticas; e *FFE-Hallu*, destinada à análise de expressões idiomáticas em persa.

Ainda que o desenvolvimento de *benchmarks* específicos tenha avançado entre 2022 e 2026, as expressões idiomáticas continuam sendo um desafio para as LLMs, mesmo diante do acesso a grandes quantidades de dados. Matheny *et al*. (2025) argumentam que, embora o treinamento específico (*fine-tuning*) contribua para melhorar o desempenho, permanece a necessidade de bases de dados mais amplas e diversificadas. Contudo, um dos grandes obstáculos atuais não reside apenas na ampliação dessas bases, mas na criação de mecanismos capazes de mensurar se as LLMs realmente compreendem o "âmago" das expressões idiomáticas ou se apenas reproduzem padrões estatísticos de probabilidade linguística.

Os esforços acadêmicos voltados ao desenvolvimento de *benchmarks* específicos — como *rolling the Dice* (Mi *et al*., 2025), *IdioTS* e *FFE-Hallu* — revelam, em certa medida, que a avaliação das LLMs baseada apenas em métricas de desempenho mostra-se insuficiente e limitada para captar a especificidade da competência fraseológica humana. Como já discutimos em outro momento neste artigo, o problema não se resolveria com o simples aumento do volume de dados de treinamento, pois a questão central não é a cobertura estatística das expressões idiomáticas, mas sim a distinção entre a mera reprodução de padrões e a efetiva compreensão do sentido figurado. Essa distinção é fundamental para a noção de "competência fraseológica artificial simulada" apresentada para este trabalho: um modelo pode alcançar alta acurácia em tarefas de identificação ou tradução de expressões idiomáticas sem que isso implique qualquer mecanismo sociocognitivo ou sociocultural equivalente ao humano. Metodologicamente, esse comportamento das IAs generativas reforça a necessidade de uma abordagem experimental que avalie qualitativamente a adequação pragmática, a sensibilidade cultural e a coerência inferencial das respostas produzidas a partir de diferentes formas de *prompting*.

8 A competência fraseológica

Há, entre os falantes de uma determinada língua, uma habilidade multifacetada que ultrapassa o simples conhecimento de unidades fraseológicas e envolve capacidades cognitivas, linguísticas e sociopragmáticas: trata-se da competência fraseológica. Ojeda e Aguiar (1999) a compreendem como uma dimensão essencial da competência comunicativa e léxica. Arnáiz, (2019) a define como a capacidade de utilizar sequências fixas de palavras de modo adequado, conferindo ao discurso um significado mais preciso e complexo. Villavicencio Simón (2023), por sua vez, sustenta que a competência fraseológica se caracteriza pelo domínio do componente fraseológico e pela habilidade de reconhecer, compreender, interpretar e produzir sentidos fraseológicos em conformidade com a adequação discursiva.

López Vázquez (2010), nesse âmbito, distingue entre habilidades fraseológicas receptivas — voltadas à identificação e decodificação de unidades em textos — e produtivas, relacionadas à reutilização adequada dessas unidades em contextos semelhantes. Para esse autor, como as unidades fraseológicas estão intrinsecamente ligadas à cultura, a competência fraseológica depende fortemente do conhecimento cultural da comunidade linguística. Saracho Arnáiz (2019) sustenta que é essencial saber como e quando empregar a unidade fraseológica, considerando o contexto social, o interlocutor e a intenção comunicativa. Além disso, destaca que a natureza peculiar dessas

unidades — marcada pela fixidez e pela idiomaticidade — pode tornar sua aquisição um processo complexo e dificultoso para estudantes de língua estrangeira.

Enquanto Saracho Arnáiz (2019) enfatiza a importância de conhecer a estrutura externa das unidades fraseológicas (UFs) no processo de desenvolvimento da competência fraseológica, Villavicencio Simón apresenta uma perspectiva divergente ao propor uma abordagem mais transversal, na qual essa competência articula não apenas aspectos externos, mas também intralinguísticos — como gramática e léxico — e extralinguísticos, relacionados à pragmática e ao discurso. No que se refere aos processos cognitivos envolvidos, pesquisas recentes ressaltam que essa habilidade abrange o armazenamento, a recuperação e o processamento dinâmico das unidades no léxico mental, não se reduzindo à simples memorização de listas de expressões.

A competência fraseológica envolve diferentes tipos de processamentos cognitivos. Entre eles, destacam-se: o processamento no léxico mental; o metafórico e de esquemas-imagem; o composicional e de analisabilidade; e o inferencial e de implicatura. O primeiro, conforme Villavicencio Simón (2023), caracteriza-se pelo dinamismo do processamento mental do léxico em um indivíduo, que desenvolve operações cognitivas capazes de atribuir sentido às UFs ao articular conhecimentos prévios com informações novas. Nesse contexto, ressalta-se a importância das redes de significados às quais as UFs se vinculam, favorecendo a recuperação rápida na memória. Além disso, evidenciam-se os processos cognitivos que abrangem tanto a capacidade de identificar e decifrar unidades em textos quanto a de reutilizá-las em contextos semelhantes.

O segundo tipo de processamento associado à competência fraseológica refere-se à metaforicidade e aos esquemas cognitivos de imagens. Bertrán (2008) defende que as unidades fraseológicas não constituem "metáforas mortas" aprendidas isoladamente, uma vez que representam formas de pensamento, sendo interpretadas e empregadas de modo ativo pelos falantes, que projetam mentalmente uma estrutura sobre o domínio de outra. Ademais, nas UFs há um componente imagético cuja motivação sincrônica possibilita recuperar o significado literal da expressão para construir e compreender sua metaforicidade.

O terceiro tipo de processamento corresponde ao composicional e à analisabilidade. Bertrán (2008) explica que a idiomaticidade pode ser analisada em diferentes graus, uma vez que muitas unidades são processadas de forma composicional, isto é, seus componentes literais podem atuar como elementos ativos na interpretação global da unidade fraseológica. Cabe destacar que o falante pode mobilizar sua competência fraseológica para produzir variantes de uma UF ou elaborar novas metáforas a partir de um mesmo modelo — o culturema — sem o risco de ser incompreendido, o que evidencia a produtividade do pensamento fraseológico.

Já o processamento inferencial e pragmático relaciona-se tanto às inferências quanto às implicaturas. As inferências mobilizam conhecimentos enciclopédicos compartilhados, além de dados implícitos do contexto, para favorecer a compreensão de uma UF. A competência fraseológica caracteriza-se, ainda, como uma habilidade complexa e multidimensional, ao integrar conhecimentos linguísticos, sociolinguísticos e pragmáticos.

A multidimensionalidade e a transversalidade da competência fraseológica residem no fato de que ela não constitui um domínio independente, mas se situa na intersecção de diversas competências, tais como: a linguística, que abarca o conhecimento morfossintático, léxico e semântico das UFs; a sociolinguística e pragmática, que envolvem a capacidade de adequar a expressão ao interlocutor, às normas sociais, aos registros de fala e à intenção comunicativa; e a

estratégica, que corresponde à habilidade de reconhecer e produzir significados fraseológicos de acordo com as exigências de adequação do discurso (Bertrán, 2008).

A competência fraseológica, portanto, deve ser compreendida como um processo sociocognitivo dinâmico que atua no léxico mental do falante, no qual o aprendizado e o uso das UFs permitem a construção de redes de significados entre palavras, a ativação de conhecimentos prévios articulados com informações novas e dados implícitos do contexto, além do envolvimento de habilidades receptivas e produtivas das unidades (Bertrán, 2008).

Bertrán (2008) argumenta que a competência fraseológica está intrinsecamente ligada à metaforicidade, entendida como forma ativa de pensamento. Os falantes interpretam e empregam UFs ao projetar mentalmente estruturas de um domínio conceitual sobre outro, apoiando-se em modelos icônicos. Nesse processo, a cultura exerce papel fundamental, pois garante a compreensão e a reprodutibilidade das unidades. O autor destaca os culturemas — símbolos extralinguísticos verbalizados — e as palavras-chave culturais, que expressam a identidade de uma comunidade, como o tereré em Mato Grosso do Sul. Além disso, as UFs transmitem simbolismos compartilhados, enraizados na coletividade por meio do conhecimento enciclopédico. A experiência psicomotora corpórea também constitui base da competência fraseológica, já que muitos esquemas emergem de atividades sensório-motoras (como "para cima é bom/para baixo é ruim"), funcionando como geradores universais de metáforas. Por fim, os somatismos — UFs que incluem partes do corpo — reforçam o caráter antropocêntrico da linguagem, sendo encontrados em todas as línguas catalogadas e apresentando significados geralmente mais transparentes.

9 Competência Fraseológica Humana *versus* Competência Fraseológica Artificial: fundamentos para uma teoria integrada

Este artigo posiciona-se teoricamente em uma abordagem sociocognitiva que articula duas grandes vertentes do pensamento linguístico contemporâneo: a Linguística Cognitiva, representada pelos trabalhos basilares de Lakoff & Johnson (1980), pelas pesquisas sobre corporalidade e simulação mental de Gibbs Jr. e Macedo (2010) e pelas contribuições de Bertrán sobre a produtividade icônica das unidades fraseológicas; e a Fraseologia de tradição eurasiática, refletidas nas obras de Corpas Pastor (1996) e Dobrovol'skij e Piirainen (2005).

A escolha desses dois eixos não é meramente intuitiva, mas considera-se como basilar tecermos o argumento central desta pesquisa, isto é, defendemos que a competência fraseológica humana desenvolve-se da intersecção e articulação entre estruturas cognitivas corporalizadas; experiências socioemocionais partilhadas e sistemas simbólico-culturais historicamente sedimentados - dimensões que, conforme tentaremos demonstrar neste trabalho - (ainda) permanecem inacessíveis aos Grandes Modelos de Linguagem em sua configuração atual.

A Linguística Cognitiva configura-se como uma disciplina que busca fornecer bases explicativas para a articulação entre os sistemas conceituais e a linguagem no cérebro e na mente de humanos. Afasta-se das abordagens linguísticas tradicionais por associar pesquisas da psicologia cognitiva, neurociência e psicologia do desenvolvimento, averiguando como a linguagem refletem-se nos processos cognitivos fundamentais (Lakoff; Johnson; 1980). Nesse contexto, a metáfora não se constitui simplesmente como um recurso poético; mas, conforme Gibbs

Jr. e Macedo (2010), integra-se à maneira como os seres humanos falam e pensam em eventos e em conceitos abstratos. Logo, o sistema conceitual é, por si só, estruturalmente metafórico.

Tanto Lakoff e Johnson (1980) quanto Gibbs Jr. e Macedo (2010) concordam que a metáfora é um processo fundamental da cognição humana, arraigada em sua experiência corpórea. A metáfora, portanto, é um esquema fundamental do pensamento, advinda de padrões recorrentes da experiência físico-corporal. Além disso, defendem que as metáforas conceituais são sistemáticas e conseguem formar correspondências entre um domínio origem (a guerra, por exemplo) a um domínio destino (ex.: ele me bombardeou de perguntas).

Entendemos que a Linguística Cognitiva compreende que o pensamento humano metafórico e a figuratividade são um elo que liga as estruturas conceituais aos nossos conhecimentos de mundo. Assim, "dar com os burros n'água" ou "vestir o paletó de madeira" não são combinações de palavras com problemas semânticos: pelo contrário, são correspondências semânticas sistemáticas dos humanos, cuja vivência social e experiência corpórea projetam-se em domínios mais abstratos. Lakoff e Johnson (1980) esclarecem que as reconhecidas expressões idiomáticas de uma língua não são meramente formas fixas sem vida, mas são sistêmicas e vivas. Expressões como "ter cartas na manga" ou "apostar todas as fichas" não se estruturam conceitualmente de modo isolado do que entendemos da realidade, mas vinculam-se a uma metáfora mais ampla: a ideia de que a "vida é um jogo de azar".

A experiência corpórea, portanto, segundo Lakoff e Johnson (1980), fornece uma chave de acesso para estruturar conceitos mais abstratos. Notamos isso em expressões tais como "estar na corda bamba" ou "estar por cima" só adquirem sentido quando a mente humana consegue processá-las por meio de uma ancoragem em nossa experiência física de equilíbrio e orientação espacial.

Gibbs Jr. e Macedo (2010) salientam que a nossa capacidade de interpretar a linguagem, especialmente a metafórica, baseia-se na habilidade de imaginar ações reais no mundo externo. Ao depararmos com um elemento em sentido figurado, criamos simulações corporificadas - uma experiência contínua, tátil e cinestésica - em que as memórias não se reduzem a apenas fatos abstratos, mas se articulam em movimentos do corpo mobilizados pela mente em contato com o mundo. Essa dimensão experiencial, diferentemente de qualquer processamento estatístico-distribucional, é justamente um processamento identificado somente em humanos. Como notamos, ele é fundamental para a consolidação da competência fraseológica.

Dobrovol'skij e Piirainen (2018) defendem a importância de considerar o caráter metafórico de determinadas unidades fraseológicas a partir de perspectivas cognitivas e culturais, por meio da chamada Teoria da Linguagem Figurativa Convencional (TLFC). Para os autores, essa abordagem mostra-se mais eficaz do que a Teoria Cognitiva da Metáfora (TCM), de base lakoffiana, ao captar nuances socioculturais. Embora a TCM seja essencial para compreender processos gerais e metáforas conceituais fundamentadas na experiência corporal, revela-se limitada quando se trata de aprofundar as especificidades culturais.

A TLFC, ao integrar perspectivas cognitivas e culturais, consegue demonstrar como as unidades fraseológicas podem ser motivadas por simbolismos e esquemas de imagens compartilhados por determinadas comunidades linguísticas. Dobrovol'skij e Piirainen (2005) definem o simbolismo como um tipo de motivação semântica fundamentada no conhecimento de determinados símbolos culturais. Diferentemente da metáfora, que se apoia na similaridade física,

o simbolismo baseia-se em convenções culturais consolidadas ao longo da história por meio de tradições religiosas, literárias, mitológicas ou folclóricas. Um exemplo é a expressão idiomática verbal "ganhar o pão", no português brasileiro, que encontra equivalente no espanhol mexicano ("*ganarse los frijoles*"). Embora ambas transmitam ideias muito próximas, os símbolos divergem: no português, utiliza-se o pão; no espanhol, o feijão.

Em relação aos esquemas de imagens, Dobrovol'skij e Piirainen (2005) partem da Teoria Cognitiva da Metáfora (TCM), contudo, apresentam ressalvas quanto à sua aplicação às expressões idiomáticas. Embora as metáforas estejam vinculadas a estruturas conceituais abstratas derivadas da experiência corporal humana — em um nível genérico e quase universal — elas se mostram insuficientes para explicar as chamadas "imagens ricas" ou imagens convencionais, isto é, aquelas carregadas de significados culturais que variam conforme a comunidade linguística. O esquema de imagem básico em diversas culturas é semelhante: A MENTE É UM RECIPIENTE e A RAIVA É UM LÍQUIDO QUENTE/SOB PRESSÃO. Todavia, a "imagem rica" no Ocidente moderno (como no Brasil) associa-se a uma panela de pressão ou a um motor a vapor, como em expressões do tipo "ele explodiu de raiva" ou "saiu soltando fumaça pelas ventas". Já na cultura tradicional chinesa, o recipiente culturalmente definido não é mecânico, mas orgânico: o fígado. Nesse contexto, a raiva é concebida como o excesso de Qi (energia/fogo) que sobe por esse órgão específico, como na expressão "动肝火", que pode significar "ficar furioso" e, em tradução literal, refere-se a "subir o fogo do fígado".

A Teoria da Linguagem Figurativa Convencional, desenvolvida por Dobrovol'skij e Piirainen (2005), e o conceito de culturema, proposto por Bertrán (2008), revelam uma afinidade teórica significativa, já que ambas buscam superar as limitações da Teoria Cognitiva da Metáfora (Lakoff; Johnson, 1980). Para os três primeiros autores, embora a TCM seja eficaz na explicação de metáforas conceituais baseadas no corpo — como em TRISTE É PARA BAIXO — ela pode falhar ao capturar certas especificidades culturais que motivam grande parte das expressões idiomáticas. Nesse contexto, o conhecimento cultural torna-se essencial para compreender por que determinadas imagens são escolhidas em detrimento de outras. Um exemplo é a expressão "estar na pica do Saci", que significa estar em uma grande enrascada: sua interpretação só é possível a partir do símbolo oriundo do folclore brasileiro.

Tanto Bertrán (2008) quanto Dobrovol'skij e Piirainen (2005) destacam o papel do sentido literal e da imagem mental no processamento das expressões idiomáticas. Para eles, a imagem não deve ser entendida como uma metáfora "morta", mas sim como um elemento ativo na construção do significado idiomático. Além disso, os três autores reconhecem a existência de metáforas compartilhadas por diversas culturas — como o "fogo" associado ao perigo, presente em expressões como "brincar com fogo" —, mas ressaltam que sua interpretação pode variar. Enquanto em muitas culturas o fogo é visto como ameaça, entre os aborígenes australianos ele é considerado um aliado e um elemento curador da terra.

No que se refere às unidades fraseológicas, embora exista uma grande variedade de termos para designá-las (Corpas Pastor, 1996), diversos teóricos — Gross (1996), Corpas Pastor (1996), Xatara (1998) e Cowie (1996) — convergem quanto às características essenciais que as delimitam. Entre elas destacam-se: a natureza pluriverbal, composta por pelo menos duas palavras; a fixação, que garante a estabilidade e cristalização de seus elementos e de seu significado; a institucionalização, pela qual as expressões são reconhecidas e consagradas pela tradição cultural; e a não composicionalidade, já que o sentido global não resulta da simples soma dos significados

individuais de seus componentes. Essas propriedades se manifestam de forma particularmente evidente nas expressões idiomáticas.

Quando voltamos especificamente às expressões idiomáticas, é importante destacar os conceitos de opacidade e transparência dessas unidades, analisados a partir do valor conotativo, da motivação e da escala de abstração. Xatara (1998) detalha essa gradação ao mostrar que as expressões idiomáticas se distribuem em diferentes níveis de abstração conforme seu valor conotativo. As fortemente conotativas (opacas) apresentam bloqueio do sentido literal de seus componentes pela realidade extralinguística, dificultando a recuperação da motivação metafórica — como em "pagar o pato". Já as fracamente conotativas (transparentes) ocorrem "quando componentes semanticamente presentes, de valor denotativo, estão associados a componentes semanticamente ausentes, de valor conotativo" (Xatara, 1998, p. 172), como em "estar de mãos atadas". A tradição fraseológica russa, por sua vez, mediada por Cowie (1996) nas línguas anglo-saxãs, discute a transparência das expressões idiomáticas sob o prisma da interpretação semântica. Quando o sentido das unidades fraseológicas pode ser deduzido a partir de seus elementos constituintes — como ocorre nas expressões idiomáticas transparentes — temos uma interpretação composicional. Já nas unidades cujo significado é pré-estabelecido e consagrado pelo uso, encontramos expressões com sentidos convencionalizados, em que a interpretação não depende da soma dos significados individuais, mas da tradição linguística e cultural.

10 Competência Fraseológica Humana *versus* Competência Fraseológica Artificial Simulada: proposta de construto comparativo

À luz dos pressupostos teóricos apresentados até o momento, propomos, nesta seção, a formalização do construto central deste artigo: a competência fraseológica artificial simulada. Para tanto, e a fim de evitar descrições meramente intuitivas, buscamos delimitar esse conceito de modo operacional e comparativo, a partir de seis dimensões analíticas. Tal postura atende às exigências metodológicas necessárias para distinguir o desempenho funcional observado nos Grandes Modelos de Linguagem dos processos sociocognitivos e socioculturais que constituem a competência fraseológica dos seres humanos.

Quadro 1: Comparativo multidimensional de competências fraseológicas (humana e artificial)

Dimensão	Competência Fraseológica Humana	Competência Fraseológica Artificial Simulada
Reconhecimento idiomático	via percepção linguística, sociocultural e contextual (Villavicencio Simón, 2023; López Vázquez, 2010)	via reconhecimento distribucional estatístico (Fornaciari *et al*., 2024)
Interpretação semântica	ancorada em experiência vivida e corporalidade (Gibbs Jr. e Macedo, 2010; Lakoff & Johnson, 2003)	dependente de frequência de co-ocorrência em corpus (Li *et al.*, 2024; Cheng *et al*., 2024)

Produção contextualmente adequada	guiada por normas sociopragmáticas interiorizadas (Saracho Arnáiz, 2019; Pamies Bertrán, 2008)	guiada por padrões de saída aprendidos, sem adequação pragmática genuína (Mi *et al*., 2025)
Ancoragem cultural experiencial	via memória autobiográfica, vivência comunitária, simbolismos e culturemas (Bertrán, 2008; Dobrovol'skij & Piirainen, 2005)	Simulação estatística sem experiência subjetiva; ausência de simbolismo culturalmente sedimentado (Hosseini *et al*., 2026; Solorio *et al*., 2026)
Processamento metafórico ativo	via corporalidade, esquemas de imagem e simulação mental (Lakoff & Johnson, 1980; Gibbs Jr. e Macedo, 2010)	via representação computacional de padrões metafóricos; sem projeção conceitual corporalizada (Cheng *et al*., 2024)
Inferência sociocultural	via conhecimento enciclopédico vivido e implicaturas pragmáticas (Bertrán, 2008; Corpas Pastor, 1996)	via estatística de coocorrência em corpus massivo; heurísticas superficiais sem inferência pragmática real (Knietaite et al., 2024; Mi *et al*., 2025)

Fonte: Elaborado pelo autor.

O Quadro 1 evidencia que a competência fraseológica humana e a competência fraseológica artificial simulada não se distinguem apenas em termos de nível, mas também pela própria natureza. Sob a perspectiva da ancoragem cultural experiencial, os Grandes Modelos de Linguagem apresentam limitações estruturais que não podem ser superadas pelo simples acréscimo de dados de treinamento. Conforme defendem Hosseini *et al*. (2026) e Solorio *et al*. (2026), as IAs generativas carecem de mecanismos de validação vinculados às comunidades linguísticas reais. Em contrapartida, os falantes humanos articulam continuamente simbolismos culturalmente sedimentados e conhecimentos enciclopédicos compartilhados, que constituem, segundo Bertrán (2008), a base da produtividade fraseológica.

As dimensões de interpretação semântica, produção contextualmente adequada, processamento metafórico ativo e inferência sociocultural revelam, ao menos em termos teóricos, que os Grandes Modelos de Linguagem apresentam um desempenho funcional aceitável em tarefas controladas. Contudo, esse desempenho se ancora em heurísticas de verossimilhança estatística, e não em processos inferenciais de caráter pragmático (Mi *et al*., 2025; Knietaite *et al*., 2024). Assim, o modelo pode alcançar elevada acurácia em *benchmarks* como Rolling the Dice (Mi *et al*., 2025) ou IdioTS, sem que isso implique na ativação genuína de qualquer mecanismo equivalente à simulação mental corporificada descrita por Gibbs Jr. e Macedo (2010).

11 Como diferentes estratégias de *prompting* afetam a interpretação de unidades fraseológicas por LLMs?

O modo como diferentes estratégias de *prompting* podem modelar a interpretação de unidades fraseológicas, especialmente as expressões idiomáticas, constitui um aspecto importante para compreender a qualidade de interação humano-máquina. Neste trabalho, partimos da hipótese de que esses modelos apresentam desempenho significativamente superior ao serem submetidos a prompts semanticamente estruturados e contextualmente orientados, uma vez que essas condições podem favorecer a ativação de padrões linguísticos mais precisos e reduzir as alucinações interpretativas.

Nesse sentido, propomos o conceito de “competência fraseológica artificial simulada”, entendido como um conjunto de capacidades funcionais presentes em Grandes Modelos de Linguagem. Essas capacidades possibilitam o (re)conhecimento, a interpretação e a (re)produção de unidades fraseológicas, derivadas das representações distribucionais estatísticas de coocorrência

lexical em corpora massivos, sem que se apoiem em processos sociocognitivos, socioculturais ou corporalizados que fundamentam a competência fraseológica humana.

Esse posicionamento definicional gera implicações para a presente pesquisa, pois, ao partirmos da hipótese de que o problema das unidades não composicionais não seja meramente quantitativo, mas de natureza cognitiva e arquitetural, a estrutura do prompt torna-se uma variável fundamental. Ela funcionará como mecanismo destinado a compensar externamente as ausências internas — corporalidade, conhecimento enciclopédico vivenciado e ancoragem cultural — na competência fraseológica artificial simulada.

12 Prompting Engineering: guiando a IA para que forneça respostas mais precisas

A Engenharia de Prompt, enquanto disciplina organizada, é bastante recente, tendo-se consolidado apenas após 2020. Contudo, suas raízes conceituais remontam a anos de pesquisas voltadas para modelos de linguagem e para o processamento da linguagem natural. Pode-se compreender a Engenharia de Prompt como uma área dedicada à elaboração de instruções de entrada (prompts) destinadas a direcionar e orientar o comportamento de uma LLM, de modo a gerar respostas mais precisas, coerentes e adequadas ao contexto. Consideramos esse processo fundamental para potencializar o desempenho dessas LLMs, propiciando respostas mais satisfatórias.

Conforme Sahoo *et al*. (2025), a Engenharia de Prompt emerge como disciplina essencial para potencializar o desempenho de modelos de linguagem de grande escala pré-treinados, uma vez que sua abordagem envolve a elaboração estratégica de instruções específicas para cada tarefa, isto é, o prompt atribuído à LLM. Dessa forma, busca-se orientar a saída do modelo sem a necessidade de modificar seus parâmetros. A relevância da Engenharia de Prompt concentra-se, sobretudo, em seu impacto na ampliação da versatilidade e da adaptabilidade de uma LLM, permitindo que essas IAs se destaquem em diferentes tarefas, aplicações e contextos.

Pang *et al*. (2025) argumenta que, devido ao desempenho cada vez mais aprimorado das LLMs em diversas tarefas relacionadas ao processamento da linguagem natural — como responder perguntas ou resumir informações específicas de determinado assunto — a elaboração de bons prompts ocupa um lugar não apenas proeminente, mas essencial para o projeto a ser executado. Em outras palavras, passa de algo secundário a constituir um ponto crucial no desenvolvimento desses processos. Ainda segundo Pang *et al*. (2025), nos métodos tradicionais de aprendizado supervisionado, o modelo é ajustado principalmente por meio de parâmetros internos. Já no caso do uso adequado de prompts, o modelo passa a depender da modificação do contexto de entrada e da forma como este é formulado. Assim, ao manipularmos o contexto inicial fornecido na entrada, direcionamos o modelo para a saída desejada, sem alterar sua arquitetura interna.

Tendo em vista isso, na próxima seção, detalharemos alguns crtiérios e algumas técnicas de elaboração de prompts eficazes, na tentativa de refinar e de otimizar a interação entre humanos e IAs.

13 Critérios para a elaboração de bons prompts: o que as pesquisas mostram

Estudos recentes têm demonstrado que a engenharia de *prompts* vem se aprimorando no sentido de estabelecer uma comunicação efetiva e eficaz entre a intenção humana e os modelos de linguagem em larga escala. *Prompts* bem elaborados possibilitam maior precisão, confiabilidade e segurança nas respostas, além de contribuírem para a redução de alucinações em diferentes domínios solicitados às LLMs.

Um bom *prompt*, em um contexto interativo entre humano e máquina, deve ser elaborado com base em quatro componentes fundamentais: uma instrução clara, um contexto relevante, dados de entrada adequados e um indicador de saída que delimite apropriadamente o formato esperado da resposta (Colangelo *et al*., 2025). Para que um *prompt* seja eficaz, a clareza e a precisão são indispensáveis, pois minimizam a possibilidade de ambiguidades, linguagem vaga e raciocínios pouco delimitados. Todos esses fatores afetam diretamente a consistência e a reprodutibilidade da resposta. Ademais, é importante ressaltar que a definição detalhada e explícita do formato e do estilo desejados — que podem incluir listas, JSON, rótulos, entre outros — propicia uma melhora significativa nos resultados (Chen *et al*., 2024).

Sivarajkumar *et al*. (2024) acrescentam três elementos complementares à elaboração de bons prompts: a especificidade, o contexto e o exemplo. A primeira é essencial, pois reduz a probabilidade de respostas alucinadas ou irrelevantes. O segundo contribui para maior acurácia e alinhamento entre a resposta e a tarefa solicitada, permitindo que o sistema opere dentro de parâmetros bem definidos. Por fim, a inclusão de exemplos, especialmente em formato few-shot prompting, mostra-se relevante em diversas tarefas, sobretudo nas mais complexas. Assim, a combinação desses três elementos complementares pode enrobustecer o prompt, tornando-o mais instrutivo e eficaz na orientação para a geração de respostas de alta qualidade. Vejamos o exemplo:

"Como especialista em Fraseologia, analise a expressão "comer barriga": explique os sentidos literal e figurado, dê três exemplos de uso e indique um equivalente em inglês. Contexto: pesquisa acadêmica sobre fraseologia. Não invente dados. Exemplo: "chutar o balde" → figurado: abandonar restrições; inglês: *to throw caution to the wind*".

Este prompt delimita a análise da expressão "comer barriga" em três dimensões claras, orienta para uma comparação intercultural (português e inglês), além de fornecer um modelo de saída (few-shot prompting).

14 Tipologia de Prompts

Os estudos recentes sobre prompting classificam-no tanto pela quantidade de informação fornecida ao modelo quanto pela estrutura de raciocínio oferecida ao LLM. Conforme Sivarajkumar *et al*. (2024), a partir de sua organização, os prompts podem ser categorizados em:

14.1 Zero-shot prompting

Nesse caso, o modelo de LLM recebe um comando sem que o usuário forneça exemplos prévios. Assim, espera-se que a resposta seja produzida exclusivamente a partir do conhecimento interno do modelo. Para Chen *et al.* (2025, p. 1, tradução nossa):

> [...]executar tarefas inéditas sem treinamento adicional ou exemplos específicos para essas tarefas é chamado de aprendizado zero-shot. Nesse contexto, os grandes modelos de linguagem (LLMs) têm demonstrado possuir uma forte capacidade zero-shot.

Logo, exigimos nessa técnica que o modelo execute uma tarefa sem ter sido treinado especificamente para ela ou recebido exemplos que o contextualizem sobre o tema.

14.2 Few-shot prompting

Neste caso, o usuário fornece ao modelo de linguagem uma instrução inserida diretamente no prompt, apresentando alguns exemplos rotulados da tarefa antes de solicitar a resposta. Contudo, não se trata de um novo treinamento, mas sim de uma forma de explorar a capacidade de *in-context learning* (aprendizagem em contexto) já presente no LLM. Nas palavras de Sivarajkumar (2024, p. 2, tradução nossa):

> Um dos aspectos fundamentais da engenharia de prompt é a quantidade de exemplos, ou shots, fornecidos ao modelo junto com o prompt. O agrupamento de prompts com poucos exemplos (*few-shot prompting*) é uma técnica que fornece ao modelo alguns exemplos de pares de entrada e saída, enquanto o agrupamento de prompts sem exemplos (*zero-shot prompting*) não fornece nenhum exemplo [...].

Desse modo, a técnica de *few-shot prompting* oferece ao modelo exemplos tanto de entrada como de saída para que ele entenda o padrão almejado. (Chen *et al*., 2023).

14.3 Chain-of-Thought (CoT*)*

Nesta modalidade de *prompting*, o usuário induz o modelo a resolver um problema complexo por meio da decomposição em etapas lógicas de raciocínio. Chen *et al.* (2023) destacam que o *Chain of Thought (CoT)* constitui uma técnica de elaboração recente no campo de estudos e aplicações de IAs e que vem demonstrando melhorias consistentes na acurácia das LLMs. Para esses autores:

> O agrupamento de *prompts* por Cadeia de Pensamento *(CoT prompting*) envolve o fornecimento de etapas intermediárias de raciocínio para guiar as respostas do modelo, o que pode ser facilitado por meio de *prompts* simples, como 'Vamos pensar passo a passo', ou por uma série de demonstrações manuais, cada uma composta por uma pergunta e uma cadeia de raciocínio que leva a uma resposta. Isso também fornece uma estrutura clara para o processo de raciocínio do modelo, tornando mais fácil para os usuários entenderem como o modelo chegou às suas conclusões (Chen *et al*., 2023, p. 6, tradução nossa).

Torna-se relevante destacar, a partir dessa leitura, que há uma solicitação de um passo a passo (*step by step*) ao modelo, a fim de uma melhor clareza e eficácia na resposta.

Cabe ainda mencionar, entre várias outras, uma última forma de prompt. Njifenjou *et al.* (2024) explicam que o *role-play prompting* consiste em uma técnica na qual o usuário interage com o LLM, ativando ou favorecendo determinados simulacros já presentes na IA. Esses múltiplos simulacros foram assimilados pelo próprio modelo durante o treinamento inicial que lhe deu origem, bem como por meio das constantes atualizações realizadas por seus criadores (como formas de falar, estilos de escrita, papéis sociais, situações comunicativas, entre outros). Os simulacros, entretanto, não são disponibilizados simultaneamente. Por isso, a LLM precisa ser direcionada quanto ao caminho a seguir, de modo a ativar o modo mais adequado.

A técnica de *role-play prompting* configura-se justamente nesse processo de direcionamento da IA, levando-a a assumir um papel, estilo ou perspectiva que favoreça uma interação mais eficaz. A LLM não se limita a interpretar uma personagem, podendo também adotar um estilo discursivo, uma situação comunicativa ou mesmo um idioma específico.

Reconhecidos todos esses pressupostos teóricos, passemos à explicitação da metodologia desta pesquisa, bem como, na sequência, à apresentação dos dados de sua análise.

15 Metodologia

Esta pesquisa configura-se como um estudo experimental de abordagem mista, uma vez que combina procedimentos quantitativos e qualitativos para investigar o desempenho de um Grande Modelo de Linguagem, Claude Haiku 4.5, no processamento de expressões idiomáticas do português brasileiro. O estudo consistiu em um experimento quase controlado, no qual se avaliou a influência de diferentes estratégias de *prompting* sobre a capacidade do modelo de reconhecer, interpretar e contextualizar unidades fraseológicas.

O desenho metodológico foi orientado pela hipótese central de que os LLMs apresentam uma competência fraseológica artificial simulada, entendida como um desempenho funcional oriundo de regularidades estatísticas aprendidas durante o treinamento, sem que isso implique a existência de mecanismos sociocognitivos, pragmáticos ou culturalmente ancorados equivalentes aos dos humanos. Nesse contexto, a variável independente correspondeu ao tipo de *prompting* empregado, enquanto a variável dependente configurou-se na qualidade da competência fraseológica artificial simulada, aferida na detecção de idiomaticidade (D1), precisão semântica (D2), adequação pragmática (D3) e sensibilidade cultural (D4).

O modelo submetido à análise foi o Claude Haiku 4.5, acessado por meio da interface conversacional claude.ai. A escolha por um único modelo justifica-se pela necessidade de controle rigoroso das condições experimentais, por se tratar de uma pesquisa piloto e pelo foco na caracterização qualitativa e quantitativa do desempenho analítico fraseológico de um sistema específico, considerando que generalizações em múltiplos modelos demandariam corpus e protocolos de escopo mais amplo.

O corpus experimental foi constituído por 140 expressões idiomáticas do português brasileiro, distribuídas em quatro grupos de 35 itens cada (EIs opacas; EIs somáticas; EIs culturais; e EIs fabricadas). O levantamento das unidades (com exceção das fabricadas) foi solicitado ao modelo Gemini 3.5 Flash, complementado pela verificação de sua convencionalidade na Web como corpus e em fontes lexicográficas (como Xatara, 2013). As 35 expressões idiomáticas propositalmente inventadas pelo pesquisador humano, denominadas neste estudo de contraprovas experimentais, foram criadas com base em padrões morfossintáticos e lexicais potencialmente plausíveis no português brasileiro (como “mandar para as cacaias”, “viajar no ketchup” ou “fazer nelorinho”). A ausência de convencionalização foi atestada por buscas sistemáticas nos corpora mencionados anteriormente. Esse procedimento teve por finalidade avaliar a capacidade do modelo de distinguir entre as expressões idiomáticas genuínas e as combinações lexicais verossímeis, mas inexistentes na língua portuguesa, o que permitiria a mensuração da incidência do fenômeno denominado neste trabalho de alucinação figurativa.

Com relação à estratificação do corpus, baseamo-nos nos seguintes critérios:

Quadro 2: Estratificação do corpus

Tipo de UF	N	Justificativa teórica
EIs opacas (p. ex., "bater as botas")	35	Elevada não composicionalidade; (Xatara, 1998; Li *et al*., 2024)

EIs culturais (p. ex., "na pica do Saci")	35	Mobilizam culturemas e simbolismos historicamente sedimentados (Bertrán, 2008; Dobrovol'skij & Piirainen, 2005)
Somatismos (p. ex., "custar os olhos da cara")	35	Caráter antropocêntrico (Bertrán, 2008)
EIs fabricadas (inexistentes)	35	Contraprova experimental para aferição da capacidade de rejeição do modelo; verificação de alucinação figurativa
Total	**140**	

Fonte: elaborado pelo autor

A opção por quatro grupos equiparados (35 UFs por grupo) visa ao equilíbrio amostral e permite comparações intergrupais controladas nas análises estatísticas. Para cada EI presente em nosso corpus, o modelo recebeu três tipos de prompt, aplicados em sessões diferentes e independentes, com histórico conversacional zerado a cada rodada, a fim de eliminar o efeito de memória contextual entre as condições. As três condições experimentais são descritas a seguir.

Na condição *zero-shot*, o modelo recebeu uma instrução direta, sem contexto adicional, exemplos prévios ou orientações quanto ao formato de resposta esperado. O formato foi: “O que significa a expressão [UF] em português brasileiro?”. Essa condição buscou replicar o padrão de interação realizada de modo espontâneo com o modelo e serviu como base para a comparação com as condições mais estruturadas. Fundamentamo-nos nas pesquisas de Huang et al. (2025) sobre as capacidades de aprendizado zero-shot em LLMs.

Na condição *few-shot* contextualizado, por sua vez, o prompt foi estruturado com a finalidade de ativar a capacidade de *in-context learning* do modelo, fornecendo-lhe, antes da apresentação da tarefa principal, três exemplos rotulados como “entrada/saída”. A estrutura do prompt incluiu: (a) atribuição de papel especializado ao modelo; (b) três exemplos de análise de EIs autênticas, bem como seus respectivos contextos de uso e explicitação de sua metaforicidade; e (c) solicitação de análise da UF almejada, com exemplo de contexto de uso em sentença. A formatação em *few-shot* visa compensar parcialmente a ausência de ancoragem contextual da condição *zero-shot,* conforme defendido por Sivarajkumar *et al*. (2024) e Chen *et al*. (2024), além de verificar em que proporção o fornecimento de exemplos modifica o desempenho do modelo nas quatro dimensões requeridas. O prompt modelo foi: “Você é um especialista em Fraseologia do Português Brasileiro. Abaixo estão exemplos de análise de expressões idiomáticas: Exemplo 1: Entrada: dormir com as galinhas. Entrada: ele dormiu com as galinhas. Saída: Expressão idiomática verbal que significa dormir muito cedo. Exemplo 2: descer o reio. Entrada: O pai desceu o reio na pobre criança. Saída: expressão idiomática que significa bater, agredir. Exemplo 3: botar chifre. Entrada: O marido botou chifre na esposa. Saída: Expressão idiomática que significa trair. Agora siga o mesmo padrão para a expressão a seguir: Expressão Idiomática: [EXEMPLO DE UF]”.

Entrada: [FRASE CRIADA PELO PESQUISADOR EM QUE SE UTILIZA A EI SOLICITADA].
Saída:

Já a condição *Chain-of-Thought* com *role-playing* constitui a estrutura mais complexa entre as três condições experimentais. O prompt apresenta: (a) atribuição de papel especializado de alto nível — linguista com doutorado em Fraseologia e especialização em Linguística Cognitiva — consolidando a técnica de *role-playing prompting* (Njifenjou *et al*., 2024); (b) protocolo de raciocínio passo a passo, em seis etapas sequencialmente descritas (*Chain-of-Thought*), abarcando identificação da UF com validação empírica, análise da estrutura lexical, demonstração de não composicionalidade, caracterização do sentido metafórico convencionalizado, descrição do contexto sociocultural de uso e apresentação de lista de evidências e referências; (c) comando de validação obrigatória por meio de corpora linguísticos de referência; e (d) critério de interrupção da análise, caso a expressão não seja reconhecida como UF convencionalizada. O prompt modelo empregado foi: "Você é um linguista com doutorado em Fraseologia e ampla especialização em Linguística Cognitiva, com foco em expressões idiomáticas do português brasileiro. Analise a expressão [EXEMPLO DE UF], seguindo o procedimento abaixo. Instruções gerais: Realize o raciocínio internamente e apresente apenas as conclusões de cada etapa. Seja técnico, preciso e utilize terminologia linguística apropriada. Baseie as conclusões em evidências empíricas verificáveis provenientes de corpora e fontes confiáveis. Priorize dados de uso real do português brasileiro obtidos em corpora linguísticos. Se a expressão não for uma expressão idiomática do português brasileiro, interrompa a análise após a Etapa 1 e justifique com evidências. Fontes prioritárias para validação: WebCorp / WebCorpus; Corpus do Português (Davies); Corpus Brasileiro (PUC-SP); Linguateca / AC/DC; Corpus NILC; Sketch Engine (quando disponível); CETEM Público (para contraste lusófono, se pertinente); Dicionários fraseológicos, lexicográficos e literatura acadêmica especializada. Etapa 1 — Identificação fraseológica e validação: Determine se a expressão é uma expressão idiomática convencionalizada no português brasileiro. Comprove com evidências de corpus (frequência, ocorrências, distribuição contextual, atestação em fontes lexicográficas ou fraseológicas). Etapa 2 — Estrutura lexical e sentido literal: Identifique os componentes lexicais da expressão e descreva o sentido composicional/literal de cada elemento e da construção completa. Etapa 3 — Não composicionalidade: Explique por que o significado literal não explica o uso real observado nos corpora. Utilize exemplos atestados para demonstrar o afastamento entre sentido literal e uso convencional. Etapa 4 — Sentido idiomático convencionalizado: Apresente o significado idiomático estabilizado e sustente a interpretação com ocorrências reais, contextos de uso e fontes especializadas. Etapa 5 — Contexto sociocultural de uso. Indique registro, grau de formalidade, distribuição sociocultural, variações regionais (quando existirem) e contextos pragmáticos observados nos dados. Etapa 6 — Evidências e referências: Liste explicitamente: Corpus consultados; Quantidade de ocorrências (quando disponível); Exemplos reais ou trechos representativos; Dicionários ou estudos utilizados; Limitações dos dados (caso existam). Formato da resposta: Status idiomático + evidências; Componentes lexicais e sentido literal; Não composicionalidade; Sentido idiomático; Contexto sociocultural; Evidências empíricas e referências; Critério obrigatório: nenhuma conclusão deve ser apresentada sem sustentação em evidências empíricas ou fontes linguísticas confiáveis".

Com relação à operacionalização das dimensões analíticas, cada resposta gerada pelo Claude Haiku 4.5 foi avaliada pelo pesquisador em quatro dimensões, com escores atribuídos em escala crescente (0 a 3), conforme descrito na tabela abaixo. O quadro apresenta as dimensões, seus critérios avaliativos, bem como os parâmetros de pontuação.

Quadro 3: Operacionalização das dimensões analíticas

Dimensão	O que avalia	Âncora teórica	Escala (0–3)
D1 — Detecção idiomática	O modelo identificou corretamente que a expressão é idiomática (não literal)?	Fornaciari *et al*. (2024); paradoxo contextual (Cheng *et al*., 2024)	0 = ausente; 1 = parcial; 2 = correto; 3 = correto e justificado
D2 — Acurácia semântica	O significado idiomático fornecido é correto e corresponde ao sentido convencionalizado?	Xatara (1998); não composicionalidade *Li et al.*, 2024.	0 = ausente; 1 = parcial; 2 = correto; 3 = correto e detalhado
D3 — Adequação pragmática	O modelo indicou contexto de uso, registro e restrições situacionais adequados?	López Vázquez (2010); Saracho Arnáiz (2019)	0 = ausente; 1 = parcial; 2 = correto; 3 = correto e contextualizado
D4 — Sensibilidade cultural	A resposta demonstra consciência da ancoragem cultural, dos culturemas e do simbolismo associado à UF?	Bertrán (2008); Dobrovol'skij & Piirainen (2005)	0 = ausente; 1 = parcial; 2 = correto; 3 = correto, historicamente situado

Fonte: Elaborado pelo autor.

As 140 unidades foram submetidas ao modelo nas três condições experimentais, gerando um total de 420 pares de estímulo-resposta. Cada condição foi aplicada em sessões independentes, com histórico conversacional zerado, a fim de mitigar qualquer influência de turnos anteriores sobre as respostas subsequentes. As respostas foram avaliadas integralmente pelo pesquisador humano e registradas em arquivos institucionais (universidade ao qual está vinculado) e pessoais. A coleta foi precedida por um estudo-piloto, no qual um subconjunto de UFs foi submetido às três condições experimentais, de modo que os resultados servissem para ajustar a formulação dos prompts, identificar possíveis ambiguidades nas instruções e verificar a viabilidade operacional do protocolo avaliativo. Cabe salientar que essa etapa contou com o auxílio do modelo ChatGPT 5.4 Fast.

Com relação à análise quantitativa, os dados foram organizados por meio de estatística descritiva, incluindo o cálculo das médias e desvios-padrão dos escores por dimensão (D1 a D4) e por subcategoria do corpus (opacas, culturais, somáticas, inexistentes). Para a comparação entre condições, foram calculadas, com auxílio do modelo DeepSeek V4 e do Kimi K2.6, as distribuições de frequência dos escores em cada dimensão e os percentuais de respostas classificadas em cada nível da escala. Já a análise qualitativa concentrou-se em três padrões de comportamento do modelo: (a) ocorrência de alucinação figurativa; (b) confusões entre expressões inexistentes e UFs semanticamente próximas; (c) diferenças qualitativas no comportamento do modelo entre as condições experimentais.

Esta pesquisa não envolveu sujeitos humanos como objeto de análise nem dados de caráter pessoal ou sensível. Os dados analisados consistem exclusivamente em respostas geradas por um sistema computacional em reação a estímulos linguísticos elaborados pelo próprio pesquisador. Com relação ao uso de inteligência artificial no processo de pesquisa, foram empregados: ChatGPT 5.4 Fast (*brainstorming* e refinamento dos prompts), Gemini 3.5 Flash (levantamento inicial de UFs, com validação posterior em corpora pelo pesquisador), Claude Haiku 4.5 (aplicação do protocolo experimental), DeepSeek V4 (organização dos dados e cálculos estatísticos) e Microsoft 365 Copilot (revisão gramatical do manuscrito) — estando declarado explicitamente nesta seção, em conformidade com as diretrizes de transparência metodológica adotadas por periódicos científicos.

16 Comparação entre condições de *prompting*

Ao compararmos as diferentes estratégias de *prompting*, é possível notar que o desempenho do modelo varia significativamente a depender do formato das instruções e a estruturação da tarefa.

As condições *few-shot* conseguiram apresentar resultados claramente superiores aos observados em zero-shot. Nas expressões idiomáticas opacas, por exemplo, a detecção de idiomaticidade (D1) alcançou desempenho satisfatório, com uma média de "3" em uma escala de 0 a 3. A precisão semântica, por sua vez, demonstrou um desempenho notável, uma vez que não houve registros de interpretações inadequadas. Entretanto, as dimensões: pragmática (D3: 90,9% concentrado na nota 2) e cultural (D4: 66,7% das respostas na nota 1 e 21,2% na nota 2) não apresentaram desempenho satisfatório, o que sugere que a melhoria de contexto possibilitada pela condição *few-shot*, em comparação com a zero-shot, concentrou-se principalmente nos aspectos semânticos.

Já quando o corpus foi submetido à condição CoT, obtivemos resultados mais aprimorados. A explicitação do raciocínio favoreceu a análise criteriosa da não composicionalidade, da validação lexical e da contextualização pragmática. Com relação ao conjunto de EIs convencionais no português brasileiro, a estratégia resultou em níveis elevados de reconhecimento e em perspectivas metalinguísticas mais detalhadas. Apesar da evidente melhora com a inclusão da explicitação do raciocínio, não foi possível solucionar completamente a tendência do modelo de produzir interpretações plausíveis para expressões inexistentes (8,5%), quando estas demonstraram elevada semelhança estrutural com unidades fraseológicas legítimas e cristalizadas no português brasileiro (como em "mandar catar boldinho").

16 Desempenho nas dimensões fraseológicas

Ao analisarmos as quatro dimensões avaliativas, percebe-se um padrão recorrente em praticamente todos os experimentos. Vejamos cada uma das dimensões mais detalhadamente (cabe salientar que, na sequência, apresentamos a média de cada dimensão para as quatro modalidades de expressões presentes em nosso corpus).

A dimensão D1 (detecção da idiomaticidade) demonstrou elevada sensibilidade em relação às características do corpus empregado nesta pesquisa. Nas parcelas de EIs autênticas, os índices de reconhecimento oscilaram entre 80% e 100%. No entanto, quando incluímos expressões inventadas, observamos uma queda expressiva na precisão classificatória. As falhas mais consistentes ocorreram em EIs inexistentes construídas a partir da semelhança com padrões recorrentes no português brasileiro, principalmente estruturas como "verbo + sintagma nominal"

(por exemplo, “passar roupa suja”). Nesses casos, a familiaridade com a estrutura da expressão existente (como “lavar roupa suja”, por exemplo) parece ter induzido o modelo a considerar como real a existência dessa unidade, sem evidências de convencionalização.

Com relação à precisão semântica (D2), esta configurou-se como a dimensão mais consistente em todos os experimentos. Em vários conjuntos de expressões autênticas, as médias oscilaram entre 2,56 e 3, muitas vezes alcançando desempenho máximo. Mesmo nos casos em que o modelo falhou em indicar explicitamente a idiomaticidade de determinada unidade, as definições permaneceram semanticamente próximas aos sentidos convencionalizados, permitindo-nos inferir o grau de metaforicidade (como em “mandar catar boldinho”: “Registro: Coloquial, Formalidade: Baixa, Distribuição sociocultural: Transversal — mas sem lastro em cultura brasileira real”). Esse padrão nos leva a considerar que o Claude possui representações semânticas desenvolvidas pelos seus treinamentos, conseguindo recuperar significados figurados, ainda que nem sempre consiga classificá-los adequadamente como uma unidade fraseológica.

Os resultados obtidos na adequação pragmática (D3) foram os mais heterogêneos, pois, embora o modelo frequentemente identificasse contextos de uso plausíveis e registros discursivos estruturalmente adequados, observamos uma tendência à superficialidade na descrição de restrições situacionais, intenções comunicativas e efeitos pragmáticos. Nos experimentos realizados, as médias permaneceram abaixo das observadas para D1 e D2, indicando desempenho moderado nessa dimensão (variando de 2,56 a 2,85 na média geral).

A dimensão cultural (D4) apresentou os resultados mais baixos em praticamente todas as parcelas do corpus analisado. Ainda que o modelo tenha demonstrado capacidade de identificar referências culturais amplamente difundidas (como “loira burra”), é possível verificar a dificuldade recorrente na explicitação de culturemas, motivações históricas e elementos socioculturais específicos relacionados às EIs. A contextualização cultural, quando presente, mostrou-se frequentemente genérica e com desempenho inferior ao das análises semânticas (“loira burra” em condiçãp *zero-shot*: “É frequentemente usada em piadas, memes e conteúdo humorístico na internet e na cultura popular brasileira. Perpetua um estereótipo de gênero e aparência física; Reflete preconceitos históricos sobre mulheres e beleza”).

Com relação aos padrões recorrentes de erro, em uma análise qualitativa, percebemos que se mantiveram relativamente estáveis entre os diferentes experimentos. O fenômeno mais recorrente é a alucinação figurativa. Nessas situações, o LLM não apenas classificou expressões inexistentes como idiomáticas, mas também produziu definições completas e plausíveis, exemplos de uso, contextos pragmáticos e explicações etimológicas sem qualquer respaldo empírico. Esse comportamento foi especialmente frequente em expressões fabricadas pelo pesquisador que imitavam estruturas produtivas presentes no repertório fraseológico do português brasileiro (como em “abotoar o paletó de graveto”: "Abotoar o paletó de graveto" é uma expressão idiomática do português brasileiro que significa morrer ou estar à beira da morte. A expressão é uma metáfora que brinca com a ideia de um "paletó de graveto" — uma alusão ao caixão (feito de madeira/graveto) que envolve o corpo do falecido. "Abotoar" o paletó seria, portanto, um eufemismo poético para o ato final da vida.”).

Além disso, observamos confusões claras entre expressões inexistentes e unidades fraseológicas semanticamente próximas. Em diversos contextos, o Claude interpretou uma expressão inventada como variante legítima de uma expressão efetivamente existente, transferindo-lhe o significado e o contexto de uso (como em “viajar no ketchup” versus “viajar na maionese”). Outro padrão recorrente consistiu na fabricação de motivações históricas ou culturais para justificar a suposta legitimidade de algumas expressões, sendo que essas construções narrativas apresentavam um grau convincente de plausibilidade interna, mas não encontravam respaldo em

registros lexicográficos ou corpus de referência (como em “falar mais que a mulher da aranha”, ao afirmar que, por motivos humorísticos e culturais, a EI se refere à aranha, em uma espécie de lenda folclórica, como um animal que “fala muito”).

Nos experimentos em condição de *CoT*, em que se incorporaram etapas explícitas de consulta a corpora, evidenciou-se uma melhora consistente no desempenho da classificação. No conjunto de expressões inexistentes, a análise assistida por corpora elevou o número de classificações corretas (92%), reduzindo substancialmente a incidência de falsos positivos. Além da melhora quantitativa, observamos uma mudança qualitativa no comportamento do modelo. Esse padrão sugere que a validação baseada em evidências externas atua como um mecanismo importante para decisões inicialmente guiadas por similaridade estrutural e plausibilidade estatística.

17 Interpretação dos resultados e implicações

Os resultados alcançados demonstram que o modelo Claude analisado possui um desempenho operacional elevado no que se refere ao reconhecimento e à interpretação de EIs convencionalizadas no português brasileiro. Contudo, salientamos que apresenta limitações significativas quando a solicitação exige validação sociocultural, diferenciação entre unidades autênticas e expressões propositalmente inventadas, ou a mobilização de conhecimentos culturais específicos. Essas observações são relevantes para nossa pesquisa, uma vez que buscamos delimitar as distinções entre competência fraseológica humana e competência fraseológica artificial simulada.

Sob o viés da Linguística Cognitiva, os dados obtidos nesta pesquisa sugerem que o modelo Claude Haiku 4.5 conseguiu reproduzir resultados semanticamente adequados sem necessariamente ter passado por processos equivalentes à corporalidade (*embodiment*), à simulação mental ou à projeção metafórica experiencial descritas por Lakoff e Johnson (2003) e Gibbs e Macedo (2010). A alta precisão semântica verificada na dimensão D2 evidencia que o sistema recupera com eficiência os significados metafóricos já amplamente representados em seus dados de treinamento. Todavia, a discrepância entre o desempenho satisfatório e as dificuldades observadas na validação da convencionalização aponta que a recuperação do significado metafórico não implica, necessariamente, a compreensão sociocognitiva da expressão.

No que se refere aos resultados alcançados com expressões inexistentes, estes merecem destaque. Quando confrontado com as unidades fraseológicas inventadas, o LLM muitas vezes atribuiu sentidos potencialmente críveis e, inclusive, construiu narrativas sobre a origem da expressão sem qualquer validação empírica. Essa situação constituiu uma manifestação clara do fenômeno de “alucinação configurativa”. Em vez de um simples erro na identificação e classificação das UFs, há uma tendência da IA generativa de completar as lacunas informacionais por meio de inferências probabilísticas baseadas em similaridade estrutural e distribucional.

No âmbito fraseológico, por sua vez, os resultados evidenciam que o Claude consegue atuar, de modo geral, de forma adequada quando é solicitado a recuperar essas unidades em seu repertório estatístico. No entanto, quando a instrução solicita ao modelo que determine se a sequência lexical foi efetivamente institucionalizada por uma comunidade linguística, as limitações aumentam: isso pode reforçar a tese de que a convencionalização das UFs não é simplesmente um fenômeno puramente linguístico, mas também sociocultural.

Outro ponto a ser salientado corresponde à dimensão cultural (D4), que apresentou desempenho claramente inferior às demais, quando analisada de forma global nos experimentos. Com exceção da condição *few-shot* (opacas), em que a média foi de 1,45, a avaliação da

sensibilidade cultural mostrou-se extremamente frágil no modelo. Em zero-shot (somático), a média do D4 atingiu o valor de 1,21, com 76,5% das respostas com notas inferiores a 1. Logo, podemos perceber que o LLM em questão carece de um protocolo sistemático para identificar culturemas. Embora o modelo tenha reconhecido algumas referências culturais amplamente difundidas, como o Saci ("na pica do Saci"), observamos dificuldade recorrente em explicitar culturemas, simbolismos históricos e motivações socioculturais associados às expressões idiomáticas. A EI "comer quieto", por exemplo, é identificada na condição *few-shot* como uma unidade fraseológica; contudo, carece de uma motivação sociocultural adequada (um indivíduo discreto e astuto, que evita a ostentação de suas conquistas, além de referir-se aos mineiros, cuja identidade cultural é marcada por traços de reserva, sagacidade e prudência nas interações sociais). Esses resultados apontam diretamente para as formulações de Dobrovol'skij e Piirainen (2005) e Bertrán (2008), segundo as quais a compreensão das UFs exige acesso aos conhecimentos compartilhados pelos falantes.

Também destacamos a diferença entre as condições experimentais de *prompts*. O desempenho superior observado nas condições *few-shot* e, principalmente, nos protocolos de *CoT* demonstra que uma estruturação mais elaborada das instruções funciona como um mecanismo compensatório para as limitações internas do LLM. Em outras palavras, o *prompting* parece oferecer pistas externas capazes de orientar processos de análise mais rigorosos (como os que se referem à identificação de culturemas), reduzindo a incidência de falsos positivos. Essas formas de instrução à máquina não eliminam as limitações identificadas, mas demonstram que parte delas pode ser mitigada por estratégias adequadas de engenharia de *prompt*.

18 Verificação das hipóteses da pesquisa

Observando os resultados obtidos nesta pesquisa, percebemos que a LLM — no caso, o Claude Haiku 4.5 — apresenta uma espécie de competência fraseológica operacional emergente, mas não uma competência fraseológica sociocognitiva equivalente à dos seres humanos. O modelo demonstrou desempenho consistente na identificação, interpretação e explicação de expressões idiomáticas autênticas, principalmente nas dimensões de detecção de idiomaticidade e precisão semântica. No entanto, as limitações relacionadas à validação da convencionalização, à sensibilidade cultural e à rejeição de expressões inexistentes apontam para o fato de que tal competência não se equipara à competência fraseológica humana, tal como descrita no referencial teórico deste trabalho. Ao contrastarmos o desempenho na análise de expressões reais em formatos mais completos de *prompting* com uma elevada incidência de falsos positivos em expressões fabricadas, somos levados a considerar que o modelo consegue manipular padrões fraseológicos registrados em seu repertório probabilístico, porém não dispõe de mecanismos equivalentes à experiência sociocultural compartilhada entre humanos, à memória pessoal de cada indivíduo e à inferência pragmática humana.

Embora esta pesquisa não tenha testado diretamente mecanismos cognitivos humanos, os dados obtidos da análise do modelo Claude Haiku 4.5 fornecem evidências indiretas em favor da hipótese de que a competência fraseológica humana se associa a processos cognitivos, socioculturais e experienciais que não estão presentes, ou não possuem mecanismos equivalentes, na LLM. Entretanto, o alto desempenho semântico alcançado pelo sistema aponta para o fato de que determinados resultados tradicionalmente associados à competência fraseológica humana podem ser simulados artificial e funcionalmente por mecanismos estatísticos. Desse modo, os resultados não sustentam uma distância absoluta entre desempenho humano e artificial, mas demonstram distinções importantes no que se refere à natureza dos processos subjacentes.

No que se refere à validação de *prompts* semanticamente estruturados e contextualmente delimitados, com a finalidade de produzir um desempenho superior em comparação a condições menos estruturadas, os resultados apontam fortemente que as condições *few-shot* e, especialmente, *Chain-of-Thought* alcançam desempenho superior ao obtido em condições *zero-shot.* Destacamos ainda a performance registrada em protocolos *CoT* aplicados às expressões inventadas, cuja taxa de rejeição correta atingiu 92%, contrastando com a quase incapacidade de discriminação observada em *zero-shot*.

Os resultados alcançados neste experimento sugerem que a simples disponibilidade de informação não resolve integralmente as limitações observadas ao longo da pesquisa. Mesmo quando o modelo produz interpretações semanticamente corretas, persistem equívocos relacionados à validação cultural, à convencionalização e à distinção entre plausibilidade estrutural e existência real das EIs. Contudo, é pertinente ressaltarmos que o progresso obtido por meio de *prompts* estruturados e validados por corpora indica que fatores metodológicos também podem exercer um papel importante. Por isso, ainda que possamos afirmar que os dados sustentam a tese de que existem limitações relevantes de arquitetura sistêmica no modelo, eles também demonstram que parte dessas impossibilidades pode ser atenuada por meio de estratégias adequadas de *prompting*.

Com relação à pergunta norteadora deste estudo — isto é, se os Grandes Modelos de Linguagem possuem uma competência fraseológica operacional emergente, mas não uma sociocognitiva plena equivalente à dos seres humanos — podemos considerar que, para esta pesquisa, que se limitou a analisar apenas o modelo Claude Haiku 4.5, a questão foi satisfatoriamente respondida. As evidências empíricas apontam que o modelo possui, sim, uma competência fraseológica funcional suficientemente robusta para o reconhecimento, interpretação e explicação de expressões idiomáticas convencionalizadas em variados contextos. No entanto, essa competência demonstra-se limitada quando a tarefa exige validação sociocultural, contextualização histórica, identificação de culturemas ou discriminação entre expressões idiomáticas genuínas do português brasileiro e combinações lexicalmente plausíveis. Logo, os resultados sustentam que a competência observada na referida LLM corresponde mais adequadamente ao conceito de competência fraseológica artificial simulada do que à competência fraseológica humana propriamente dita.

19 Considerações finais

Este estudo — que se limitou a analisar apenas um modelo de LLM, o Claude Haiku 4.5, com o objetivo de identificar se ele possui uma competência fraseológica artificial simulada equivalente à dos humanos — possibilitou-nos encontrar algumas contribuições teóricas e metodológicas, além de evidenciar certas limitações do modelo.

No âmbito da ciência fraseológica, os resultados reforçam a relevância de considerar a convencionalização como dimensão essencial da competência fraseológica, uma vez que a recuperação correta apenas de significados metafóricos não garante o reconhecimento adequado da legitimidade fraseológica das unidades. Para a Linguística Cognitiva, os dados apontam para posicionamentos compatíveis com estudos que demonstram a importância da corporalidade, da simulação mental e da experiência cultural na interpretação do sentido figurado. O desempenho limitado do modelo em tarefas que exigem identificação de simbolismos culturais e validação de contextos pragmáticos sugere que a competência fraseológica humana continua fortemente associada a mecanismos experienciais que ainda não encontram equivalentes diretos nos sistemas

atuais. Já para a Linguística Computacional, este estudo propõe um enquadramento teórico inovador ao formalizar o conceito de competência fraseológica artificial simulada.

Com relação às implicações metodológicas e práticas, defendemos que avaliações baseadas apenas na acurácia semântica podem superestimar as capacidades fraseológicas dos modelos, uma vez que a incorporação de dimensões relacionadas à convencionalização, à adequação pragmática e à sensibilidade cultural torna-se essencial, como apontado neste estudo. Ressaltamos ainda a importância de empregar contraprovas experimentais compostas por expressões inexistentes, pois sem esse tipo de procedimento, a capacidade do modelo de identificar idiomaticidade pode mascarar fragilidades significativas relacionadas à validação do conhecimento fraseológico. Também destacamos a relevância da validação por corpora e por fontes lexicográficas externas, já que os avanços observados após a solicitação de etapas de verificação empírica indicam que arquiteturas híbridas, capazes de combinar inteligência artificial generativa com mecanismos de consulta documental, podem mitigar significativamente a incidência de alucinações figurativas.

Essas implicações, em um plano aplicado, sugerem cautela no uso de LLMs sem supervisão humana tanto para ensino de UFs, tradução automática, elaboração de materiais didáticos e análise intercultural, porque, embora os modelos consigam oferecer explicações semanticamente plausíveis, suas respostas podem ser sistematicamente falíveis, mesmo que seja aparentemente convincente e linguisticamente adequada

Este estudo, como já destacamos, apresenta também algumas limitações que merecem ser ressaltadas. A primeira refere-se ao foco em um único modelo de LLM, pois, embora os resultados sejam consistentes para o objetivo proposto, não é possível generalizá-los automaticamente para todos os modelos atuais. A segunda corresponde a que para cada unidade fraseológica, realizamos somente uma tentativa, não replicando a análise mais de uma vez; a terceira corresponde ao recorte linguístico adotado, já que a investigação se concentrou no português brasileiro. Pesquisas futuras poderão dedicar-se a analisar: (i) se padrões semelhantes emergem em outras línguas; (ii) comparações sistemáticas entre modelos fechados e de código aberto; (iii) os papéis dos culturemas regionais com maior profundidade; (iv) o desenvolvimento de *benchmarks* especificamente voltados para a avaliação da competência fraseológica artificial simulada; e (v) a construção de protocolos que integrem a avaliação qualitativa da adequação pragmática, da sensibilidade sociocultural e da coerência inferencial, superando métricas predominantemente quantitativas.

Declaração de conflito de interesses

O autor declara que não possui conflitos de interesses financeiros, profissionais ou pessoais que possam ter influenciado a elaboração, a análise ou a interpretação dos resultados apresentados neste manuscrito.

Disponibilidade dos dados de pesquisa

Os dados de pesquisa que fundamentam os resultados deste estudo não serão disponibilizados publicamente, mas poderão visualizados mediante solicitação razoável ao autor correspondente, observadas as restrições éticas, legais e institucionais aplicáveis. Ademais, cabe salientar que os referidos dados de pesquisa que deram origem a este artigo estão armazenados em um banco de dados institucional, vinculado à Universidade Federal de Mato Grosso do Sul (UFMS)

Referências Bibliográficas

ARNÁIZ, Marta Saracho. Cómo desarrollar la competencia fraseológica en la clase de ELE. In: **LA FORMACIÓN y competencias del profesorado de ELE**: XXVI Congreso Internacional ASELE. Asociación para la Enseñanza del Español como Lengua Extranjera – ASELE, 2016. p. 921-931.

ATTIA, M. *et al*. Beyond Understanding: Evaluating the Pragmatic Gap in LLMs' Cultural Processing of Figurative Language. **arXiv preprint arXiv:2510.23828**, 2026. Disponível em: https://doi.org/10.48550/arXiv.2510.23828. Acesso em: 14 jun. 2026.

BERTRÁN, A. P. Productividad fraseológica y competencia metafórica (inter)cultural. **Language Design: Journal of Theoretical and Experimental Linguistics**, v. 10, p. 89-100, 2008.

CHEN, P. *et al*. Induce Prompting: Multi-Rationale Induction for Zero-Shot Reasoning. *In*: INUI, Kentaro *et al*. (ed.). **Findings of the Association for Computational Linguistics**: IJCNLP 2025. [*S. l.*]: Association for Computational Linguistics, 2025. Disponível em: https://aclanthology.org/2025.findings-ijcnlp.6/. Acesso em: 14 jun. 2026.

CHENG, K.; BHAT, S. No Context Needed: Contextual Quandary In Idiomatic Reasoning With Pre-Trained Language Models. *In*: Conference of the north american chapter of the association for computational linguistics: human language technologies, 2024. **Proceedings [...]** Volume 1: Long Papers. [*S. l.*]: Association for Computational Linguistics, 2024. p. 4863–4880. Disponível em: https://doi.org/10.18653/v1/2024.naacl-long.272. Acesso em: 14 jun. 2026.

COLANGELO, Maria Teresa *et al*. How to Write Effective Prompts for Screening Biomedical Literature Using Large Language Models. **BioMedInformatics**, v. 5, n. 1, p. 15-32, 2025. Disponível em: https://doi.org/10.3390/biomedinformatics5010015. Acesso em: 14 jun. 2026.

CORPAS PASTOR, Gloria. **Manual de fraseología española**. Madrid: Editorial Gredos, 1996.

COWIE, Anthony Paul. The phraseological approach. *In*: COWIE, Anthony Paul (ed.). **Phraseology**: theory, analysis, and applications. Oxford: Oxford University Press, 1998. p. 1-20.

DOBROVOL'SKIJ, D.; PIIRAINEN, E. Conventional Figurative Language Theory and idiom motivation. **Yearbook of Phraseology**, v. 9, n. 1, p. 5–30, 2018. Disponível em: https://doi.org/10.1515/phras-2018-0003. Acesso em: 14 jun. 2026.

FORNACIARI, F. D. L. *et al*. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. **arXiv preprint arXiv:2405.10579**, 2024. Disponível em: https://doi.org/10.48550/arXiv.2405.10579. Acesso em: 14 jun. 2026.

GIBBS JR, R. W.; MACEDO, A. C. P. S. D. Metaphor and embodied cognition. **DELTA: Documentação de Estudos Em Lingüística Teórica e Aplicada**, v. 26, n. esp., p. 679–700, 2010. Disponível em: https://doi.org/10.1590/S0102-44502010000300014. Acesso em: 14 jun. 2026.

GROSS, Gaston. **Les expressions figées en français**: noms composés et autres locutions. Paris: Ophrys, 1996. (Collection l'Essentiel français).

HOSSEINI, F.; YOUSEFZADEH, M.; YAGHOOBZADEH, Y. FFE-HALLU: Hallucinations in Fixed Figurative Expressions Benchmark of Idioms and Proverbs in the Persian Language. **arXiv preprint arXiv:2411.02340**, 2024. Disponível em: https://arxiv.org/abs/2411.02340. Acesso em: 14 jun. 2026.

KHOSHTAB, P. *et al*. Comparative Study of Multilingual Idioms and Similes in Large Language Models. **arXiv preprint arXiv:2410.16461**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2410.16461. Acesso em: 14 jun. 2026.

KIM, J. *et al*. Memorization or Reasoning? Exploring the Idiom Understanding of LLMs. **arXiv preprint arXiv:2505.16216**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2505.16216. Acesso em: 14 jun. 2026.

KNIETAITE, A. *et al*. Is Less More? Quality, Quantity and Context in Idiom Processing with Natural Language Models. **arXiv preprint arXiv:2405.08497**, 2024. Disponível em: https://doi.org/10.48550/arXiv.2405.08497. Acesso em: 14 jun. 2026.

LAKOFF, G.; JOHNSON, M. **Metaphors we live by**. Chicago: University of Chicago Press, 1980.

LI, S. *et al*. Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *In*: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 38., 2024. **Proceedings [...]** [*S. l.*]: AAAI Press, 2024. v. 38, n. 17, p. 18554–18563. Disponível em: https://doi.org/10.1609/aaai.v38i17.29817. Acesso em: 14 jun. 2026.

LÓPEZ VÁZQUEZ, Lucía. La competencia fraseológica en los textos de los manuales de ELE de nivel superior. *In*: SANTIAGO GUERVÓS, J. de *et al*. (coord.). **Del texto a la lengua**: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE. Salamanca: ASELE/Centro Virtual Cervantes, 2010. p. 531-542.

MI, M.; VILLAVICENCIO, A.; MOOSAVI, N. S. Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context. **arXiv preprint arXiv:2410.16069**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2410.16069. Acesso em: 14 jun. 2026.

NJIFENJOU, Ahmed *et al*. Role-Play Zero-Shot Prompting with Large Language Models for Open-Domain Human-Machine Conversation. **arXiv preprint arXiv:2406.18460**, 2024. Disponível em: https://arxiv.org/abs/2406.18460. Acesso em: 14 jun. 2026.

OJEDA, G. O.; AGUIAR, M. I. G. La competencia fraseológica y paremiológica de los hablantes canarios. *In*: PAMIES BERTRÁN, A.; LUQUE DURÁN, J. de D. (ed.). **Trabajos de fraseología y fraseodidáctica**. Frankfurt am Main: Peter Lang, 2014. p. 251-268.

PANG, Jinlong *et al*. Token Cleaning: Fine-Grained Data Selection for LLM Supervised Fine-Tuning. **arXiv preprint arXiv:2502.01968**, 2025. Disponível em: https://doi.org/10.48550/arxiv.2502.01968. Acesso em: 14 jun. 2026

REZAEIMANESH, S. *et al.* Large language models for Persian-English idiom translation. In: **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the**

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025. p. 7974-7985.

VILLAVICENCIO SIMÓN, Y. La competencia fraseológica en la enseñanza del español como lengua extranjera: Reinterpretación teórica y metodológica. **Pragmalinguistica**, n. 31, p. 483-505, 2023. Disponível em: https://doi.org/10.25267/Pragmalinguistica.2023.i31.26. Acesso em: 14 jun. 2026.

SAHOO, Pranab *et al*. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. **arXiv preprint arXiv:2402.07927**, 2025. Disponível em: https://arxiv.org/abs/2402.07927. Acesso em: 14 jun. 2026

SIVARAJKUMAR, Sonish *et al*. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. **JMIR Medical Informatics**, v. 12, p. e55318, 2024. Disponível em: https://doi.org/10.2196/55318. Acesso em: 14 jun. 2026

XATARA, C. M. Tipologia das expressões idiomáticas. **ALFA: Revista de Linguística**, São Paulo, v. 42, n. 1, 2001. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/view/4274. Acesso em: 14 jun. 2026.

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.