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: 2026-08-14 (versão 2)
(AAAA-MM-DD)

Justificativa da versão: Foram revisados detalhadamente: os dados da pesquisa, a métrica, ampliação do resumo e inserção na fundamentação teórica.

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

Introdução: Os Grandes Modelos de Linguagem (LLMs) têm demonstrado desempenho expressivo em tarefas de Processamento de Linguagem Natural, mas o processamento de unidades fraseológicas, especialmente de expressões idiomáticas (EIs), permanece teoricamente desafiador. A natureza pluriverbal, cristalizada e frequentemente não composicional dessas unidades exige não apenas o reconhecimento de padrões lexicais, mas também a mobilização de conhecimentos semânticos, pragmáticos, culturais e sociocognitivos. Nesse contexto, este estudo parte da distinção entre a competência fraseológica humana, ancorada na experiência corporal, na vivência sociocultural e na inferência pragmática, e o desempenho funcional de sistemas artificiais baseado predominantemente em regularidades estatístico-distribucionais. Objetivo: O objetivo é investigar se um LLM apresenta uma competência fraseológica operacional emergente no processamento de expressões idiomáticas do português brasileiro e verificar em que medida diferentes estratégias de prompting modificam esse desempenho. Propõe-se, para isso, o conceito de competência fraseológica artificial simulada, entendido como a capacidade funcional de reconhecer, interpretar e produzir unidades fraseológicas sem que isso implique a existência de mecanismos sociocognitivos equivalentes aos humanos. Metodologia: Foi realizado um experimento quase controlado, de abordagem mista, com o modelo Claude Haiku 4.5. O corpus experimental reuniu 140 unidades, distribuídas igualmente entre expressões idiomáticas opacas, somatismos, expressões culturalmente marcadas e expressões fabricadas como contraprovas experimentais. Cada unidade foi submetida a três condições independentes de prompting: zero-shot, few-shot contextualizado e Chain-of-Thought associado a role-playing. O desempenho foi analisado em quatro dimensões: detecção da idiomaticidade (D1), precisão semântica (D2), adequação pragmática (D3) e sensibilidade cultural (D4). Resultados: Os resultados indicam elevada precisão semântica nas expressões convencionalizadas. A dimensão cultural revelou-se a mais frágil, sobretudo na identificação de culturemas, simbolismos e motivações socioculturais. Observou-se ainda ocorrência recorrente de alucinação figurativa, caracterizada pela aceitação de expressões inexistentes como unidades fraseológicas convencionais. Em contraste, as estratégias few-shot e, sobretudo, Chain-of-Thought produziram melhorias qualitativas, com a condição CoT atingindo alto grau de rejeição correta das expressões fabricadas. Conclusão: Os resultados sustentam a hipótese de que o modelo apresenta uma competência fraseológica operacional emergente, mas não uma competência fraseológica sociocognitiva equivalente à humana. O sistema consegue simular

determinados resultados associados à competência fraseológica por meio de padrões estatísticos, mas permanece vulnerável a falhas de convencionalização, inferência pragmática e ancoragem cultural. Ao mesmo tempo, os resultados demonstram que a engenharia de prompt pode funcionar como mecanismo compensatório parcial, sobretudo quando incorpora exemplos, contextualização, validação empírica e critérios explícitos de rejeição.

PALAVRAS-CHAVE
Fraseologia. Competência Fraseológica. Linguística Computacional. Expressões Idiomáticas. Large Language Models.

ABSTRACT
Abstract

Introduction: Large Language Models (LLMs) have achieved substantial advances in Natural Language Processing, yet the processing of phraseological units, particularly idiomatic expressions (IEs), remains theoretically challenging. Their multiword, conventionalised and frequently non-compositional nature requires more than the recognition of lexical patterns, as successful interpretation may involve semantic, pragmatic, cultural and sociocognitive knowledge. This study therefore distinguishes between human phraseological competence, grounded in embodied experience, sociocultural participation and pragmatic inference, and the functional performance of artificial systems, which is primarily based on statistical-distributional regularities. Objective: The study aims to investigate whether a large language model exhibits an emergent operational form of phraseological competence when processing Brazilian Portuguese idiomatic expressions and to determine the extent to which different prompting strategies affect its performance. To this end, the paper proposes the concept of simulated artificial phraseological competence, understood as the functional capacity to recognise, interpret and reproduce phraseological units without implying the presence of sociocognitive mechanisms equivalent to those underlying human phraseological competence. Methodology: A mixed-method, quasi-controlled experiment was conducted using Claude Haiku 4.5. The experimental corpus comprised 140 Brazilian Portuguese phraseological units, equally distributed across opaque idioms, somatic expressions, culturally marked idioms and fabricated expressions designed as experimental counter-evidence. Each unit was submitted to three independent prompting conditions: zero-shot, contextualised few-shot, and Chain-of-Thought combined with role-playing. Performance was assessed across four dimensions: idiomaticity detection (D1), semantic accuracy (D2), pragmatic appropriateness (D3), and cultural sensitivity (D4). Results: The findings indicate high semantic accuracy for conventionalised expressions, with mean scores ranging from 2.56 to 3 across different subsets. Pragmatic appropriateness was more heterogeneous, whereas cultural sensitivity consistently represented the weakest dimension, particularly in the identification of culturemes, cultural symbolism and sociocultural motivations. The analysis also revealed recurrent figurative hallucination, whereby non-existent expressions were accepted and explained as genuine idiomatic units. By contrast, few-shot prompting and, particularly, Chain-of-Thought produced qualitative improvements. The CoT condition achieved a with a high degree of correct rejection rate for fabricated expressions. Conclusion: The findings support the hypothesis that the model exhibits an emergent operational form of phraseological competence, but not a sociocognitive competence equivalent to that of human speakers. The system can simulate certain outcomes traditionally associated with phraseological competence through statistical patterning, yet remains vulnerable to failures

involving conventionalisation, pragmatic inference and cultural anchoring. At the same time, the results demonstrate that prompt engineering can partially compensate for these limitations, particularly when prompts incorporate examples, contextualisation, empirical validation and explicit rejection criteria.

KEYWORDS

Phraseology. Phraseological Competence. Computational Linguistic. Idiomatic Expression. 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), por sua vez, 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.

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 só resulta da simples soma de seus elementos. Um exemplo é "bater as botas", que significa "morrer", sem relação direta com o sentido literal das palavras. Esse tipo de construção representa um desafio para as inteligências artificiais generativas.

Pesquisas recentes, como Li et al. (2024), mostram que as LLMs frequentemente falham em captar o sentido não composicional das expressões idiomáticas, produzindo interpretações literais e traduções inadequadas. De modo complementar, Cheng e Bhat (2024) demonstram que esses modelos enfrentam dificuldades justamente em razão da natureza não composicional dessas unidades. Em conjunto, esses estudos indicam que tais falhas decorrem menos da ausência de dados nos corpora de treinamento do que do processamento distribucional das LLMs, baseado em regularidades estatísticas de coocorrência lexical, e não na projeção metafórica característica da cognição humana. Assim, a limitação não é apenas quantitativa, mas cognitiva e arquitetural: as LLMs carecem da corporalidade (embodiment) e da simulação mental (simulação corporificada).

Conforme Gibbs Jr. e Macedo (2010), a corporalidade vincula a cognição às experiências corporais, permitindo a formação de conceitos abstratos, enquanto a simulação mental possibilita compreender linguagem e metáforas por meio da imaginação de ações e da ativação de experiências sensório-motoras. Esse intervalo fundamenta o que este trabalho denomina competência fraseológica artificial simulada: um desempenho funcional que reproduz resultados da competência fraseológica humana sem os processos sociocognitivos que a sustentam.

Tanto a corporalidade quanto a simulação mental permitem que os humanos construam e atualizem a metaforicidade dos conceitos a partir de experiências sensório-motoras e socioculturais. Esse intervalo define o que chamamos de

"competência fraseológica artificial simulada": um desempenho funcional que imita a competência humana sem se apoiar nos processos sociocognitivos que a estruturam. Metodologicamente, isso reforça a necessidade de elaborar prompts que explicitem os domínios metafóricos das expressões idiomáticas, compensando a ausência de ancoragem experiencial nos LLMs.

3 Falhas no uso de Contexto: o paradoxo contextual

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 (Fornaciari et al., 2024).

5 A alucinação figurativa e as questões 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.

6 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.

A questão 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.

7 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.

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.

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

Este artigo adota uma perspectiva sociocognitiva que combina duas tradições centrais da linguística. Da Linguística Cognitiva (Lakoff & Johnson, 1980; Gibbs Jr. & Macedo, 2010; Pamies Bertrán, 2008) vem a descrição dos processos que sustentam a competência fraseológica humana; da Fraseologia eurasiática (Corpas Pastor, 1996; Dobrovol'skij & Piirainen, 2005), o instrumental para distinguir o figurado universal do dependente de validação comunitária.

Defendemos que a competência fraseológica resulta da articulação entre estruturas cognitivas corporalizadas, experiências socioemocionais e sistemas simbólico-culturais — dimensões ainda inacessíveis às LLMs. A Linguística Cognitiva mostra que a metáfora é estrutural ao pensamento humano, derivada da experiência corpórea e sistematicamente projetada em domínios abstratos. Expressões idiomáticas

como “ter cartas na manga” ou “estar na corda bamba” exemplificam como vivências físicas e sociais estruturam conceitos mais complexos. Gibbs Jr. e Macedo (2010) destacam que interpretar metáforas envolve simulações corporificadas, processo exclusivo dos humanos.

Já a Teoria da Linguagem Figurativa Convencional (TLFC), de Dobrovol’skij e Piirainen, amplia a análise ao incorporar dimensões culturais. Diferente da Teoria Cognitiva da Metáfora (TCM), que explica metáforas universais, a TLFC mostra como símbolos e imagens ricas variam entre comunidades. Exemplos incluem “ganhar o pão” no português e “ganarse los frijoles” no espanhol, ou a associação da raiva a uma panela de pressão no Ocidente *versus* ao fígado na tradição chinesa.

Assim, metáforas universais convivem com culturemas específicos, como “estar na pica do Saci”, cuja interpretação depende do folclore brasileiro. A TLFC e o conceito de culturema (Bertrán, 2008) revelam que a convencionalização cultural é decisiva para compreender expressões idiomáticas. Entre suas propriedades essenciais, destacam-se pluriverbalidade, fixação, institucionalização e não composicionalidade, além da opacidade e transparência semântica (Xatara, 1998).

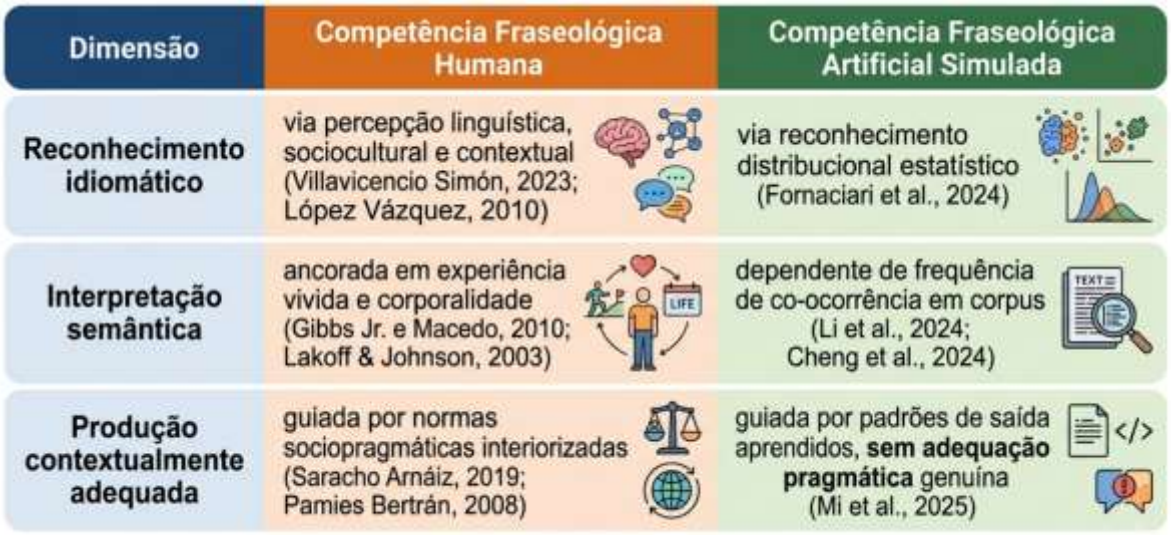
9 Competência Fraseológica Humana versus Competência Fraseológica Artificial Simulada: proposta de construto comparativo

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.

A formalização proposta dialoga com uma discussão já consolidada sobre modelos de linguagem. Mahowald et al. (2024) distinguem entre competência

linguística formal (domínio das regularidades de uma língua) e competência funcional (uso da linguagem para raciocinar e agir socialmente). Os LLMs exibem a primeira com eficiência, mas apenas parcialmente a segunda. Essa distinção retoma a objeção de Bender e Koller (2020), segundo a qual sistemas treinados apenas na forma não têm acesso pleno ao significado, e converge com a advertência de Bender et al. (2021) sobre o risco de confundir fluência com compreensão. Nosso construto, portanto, não substitui essas formulações, mas as especifica para o domínio da Fraseologia. Aqui, o critério decisivo não é composicional, mas sociocultural, isto é, uma sequência pode ser bem estruturada e interpretável, mas não existir se não for sancionada pelo uso de uma comunidade linguística. É justamente esse filtro que o modelo não aplica, como mostram os índices de alucinação figurativa. Chamamos de competência fraseológica artificial simulada a competência formal que se estende ao metafórico convencionalizado sem dispor do mecanismo comunitário de validação que, na competência humana, define o que conta como unidade da língua.

Imagem 1: Comparativo conceitual entre competência fraseológica humana e competência fraseológica artificial



Fonte: Elaborado pelo autor com auxílio do Gemini AI

A Imagem 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. 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).

10 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 (Sahoo et al., 2025).

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

Estudos recentes mostram que a engenharia de prompts busca aprimorar a comunicação entre intenção humana e LLMs, aumentando precisão, confiabilidade e reduzindo alucinações. Um prompt eficaz deve conter quatro componentes básicos: instrução clara, contexto relevante, dados de entrada adequados e indicação de saída que defina o formato esperado (Colangelo et al., 2025). Clareza e precisão são indispensáveis para evitar ambiguidades e garantir consistência. A definição explícita do estilo ou formato — listas, JSON, rótulos — também melhora os resultados (Chen et al., 2024).

Sivarajkumar et al. (2024) acrescentam três elementos complementares: especificidade, que reduz respostas irrelevantes; contexto, que aumenta acurácia e alinhamento; e exemplos, especialmente em few-shot prompting, úteis em tarefas

complexas. A combinação desses fatores torna o prompt mais instrutivo e eficaz na 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).

12 Tipologia de Prompts

Os estudos recentes sobre prompting classificam-no pela quantidade de informação fornecida ao modelo e pela estrutura de raciocínio proposta. Segundo Sivarajkumar et al. (2024), há três modos recorrentes: zero-shot, em que o LLM recebe apenas o comando; few-shot, em que o usuário insere exemplos rotulados antes da resposta; e Chain-of-Thought (CoT), que induz o modelo a resolver problemas passo a passo por meio da decomposição lógica.

Njifenjou et al. (2024) acrescentam o role-play prompting, técnica em que o usuário ativa simulacros já presentes na IA — estilos de fala, papéis sociais, situações comunicativas — assimilados no treinamento e nas atualizações. Esses simulacros não são acionados simultaneamente, exigindo que o LLM seja direcionado para ativar o modo mais adequado.

13 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).

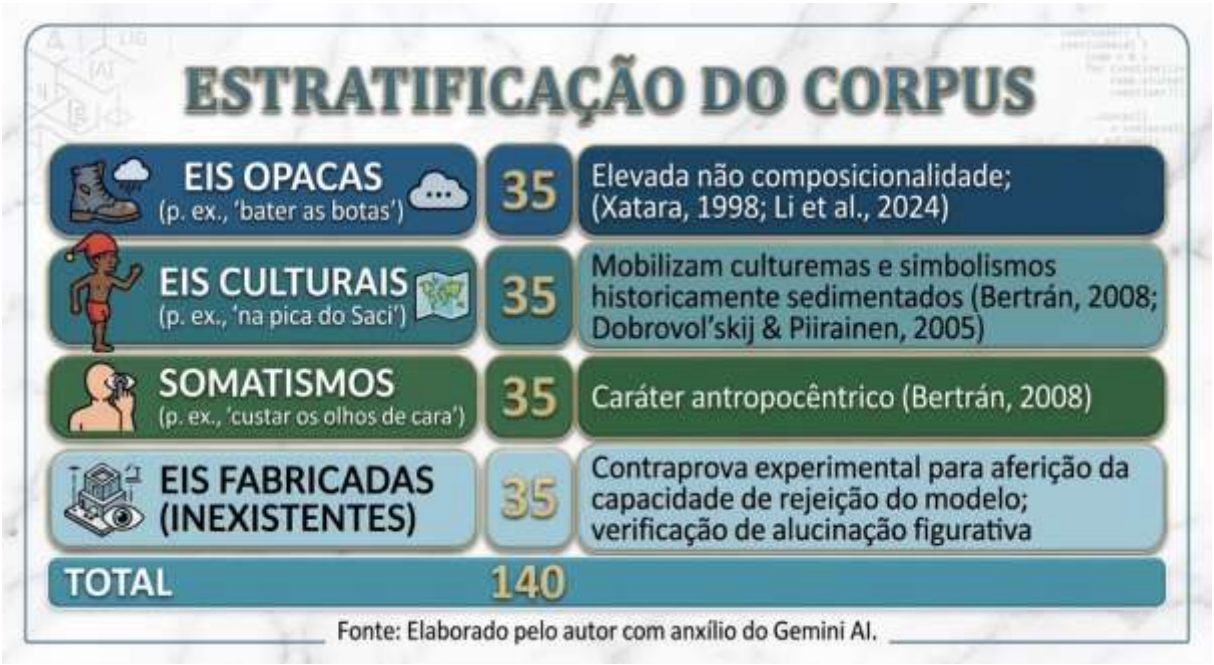
O modelo submetido à análise foi o Claude Haiku 4.5, acessado por meio da interface conversacional claude.ai. A escolha por um único modelo justifica-se pela necessidade de controle rigoroso das condições experimentais, por se tratar de uma pesquisa piloto e pelo foco na caracterização qualitativa e 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:

Imagem 2: Estratificação do Corpus



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. A imagem apresenta as dimensões, seus critérios avaliativos, bem como os parâmetros de pontuação.

Imagem 3: Dimensões de análise

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 correspond 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 com auxílio do Gemini AI

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 ChatGPT 5.4 Fast, 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, refinamento dos prompts, organização dos dados e cálculos estatísticos), Gemini 3.5 Flash (levantamento inicial de UFs, com validação posterior em corpora pelo pesquisador), Claude Haiku 4.5 (aplicação do protocolo experimental), 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.

14 Comparação entre condições de prompting

A comparação entre estratégias de prompting mostra que o desempenho do modelo varia conforme o formato das instruções e a estrutura da tarefa. Na condição few-shot, todas as expressões idiomáticas opacas tiveram sua idiomaticidade corretamente identificada (35/35), revelando desempenho máximo. Nesse cenário, as contraprovas tornam-se indispensáveis, pois só elas introduzem variação onde as unidades autênticas já não produzem.

Na condição Chain-of-Thought (CoT), os resultados foram mais refinados: a explicitação do raciocínio favoreceu a análise da não composicionalidade, da validação lexical e da contextualização pragmática, garantindo reconhecimento consistente das EIs convencionais do português brasileiro. Contudo, o método não eliminou a tendência do modelo a validar sequências inexistentes. Dois dos 24 itens fabricados (8,3%) foram aceitos como genuínos, ambos estruturalmente próximos a unidades cristalizadas, como “mandar catar boldinho”.

Embora numericamente reduzido, esse dado é teoricamente relevante: mostra que protocolos mais estruturados deslocam o limiar de aceitação do modelo, mas não estabelecem um critério efetivo de convencionalidade.

15 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. A detecção de idiomaticidade mostrou-se fortemente sensível à natureza do item avaliado. Nas unidades convencionalizadas, o reconhecimento foi notoriamente alto nas três condições, conforme os valores discriminados na Tabela 1, que apresenta as taxas por categoria e por condição com os respectivos denominadores. A queda expressiva

ocorre nas contraprovas experimentais e concentra-se, de modo sistemático, nos itens construídos por analogia com padrões produtivos do português brasileiro, sobretudo os de estrutura verbo mais sintagma nominal. É o que se observa em "passar roupa suja", cuja proximidade formal com a unidade cristalizada "lavar roupa suja" parece ter induzido o modelo a pressupor a existência da sequência sem qualquer evidência de convencionalização. Optamos por reportar os valores célula a célula, e não por faixas agregadas, porque as taxas variam substancialmente conforme a categoria e a condição, de modo que um intervalo global obscureceria justamente a assimetria que constitui nosso objeto.

Com relação à precisão semântica (D2), esta configurou-se como a dimensão mais consistente em todos os experimentos. A acurácia semântica foi a dimensão mais estável do experimento, e essa estabilidade verifica-se nas três condições e em todas as categorias de unidades convencionalizadas, conforme os valores discriminados na Tabela 1. Mesmo nos casos em que o modelo falhou em declarar explicitamente o caráter idiomático da unidade, as definições permaneceram próximas do sentido convencionalizado, comportamento que atribuímos à robustez das representações semânticas distribucionais adquiridas durante o pré-treinamento. O ponto que nos interessa reter é que essa robustez semântica convive com fragilidade acentuada na validação da convencionalização, e é precisamente essa dissociação — e não o desempenho semântico tomado isoladamente — que sustenta o construto proposto neste trabalho.

A adequação pragmática (D3), por sua vez, demonstrou um desempenho mediano: 90,9% das respostas concentraram-se na nota 2 da escala, o que corresponde a 30 dos 33 itens efetivamente avaliados nessa célula e descreve indicações de uso corretas, porém não contextualizadas. Registre-se que dois dos trinta e cinco itens do grupo não puderam ser avaliados nessa dimensão, em razão de falhas no armazenamento digital feito pelo pesquisador, de modo que os percentuais aqui reportados tomam como denominador os 33 itens analisados.

A sensibilidade cultural (D4) foi, também nessa condição, a dimensão de pior desempenho. Das 33 respostas avaliadas, 66,7% receberam nota 1, 21,2% receberam nota 2 e 12,1% alcançaram nota 3, sem qualquer ocorrência de nota 0 (distribuição que resulta em média de 1,45). Em outras palavras, mesmo na configuração experimental em que o modelo obteve seu melhor resultado com relação ao cultural, dois terços das respostas limitaram-se a um reconhecimento apenas parcial da ancoragem sociocultural da unidade.

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").

Na condição CoT, que exigiu evidência empírica de corpora de referência, observou-se melhora consistente na classificação das contraprovas. No entanto, cabe-se uma ressalva: o Claude Haiku 4.5, acessado pela interface conversacional, não dispõe de acesso efetivo aos corpora mencionados no prompt. Assim, as referências a frequências, ocorrências e atestações produzidas pelo modelo não resultam de consultas documentais concomitantes, mas reproduzem discursivamente a forma da evidência de corpus. O ganho observado não decorre, portanto, de verificação empírica externa. Interpretamos a exigência de justificação empírica como um um modo que "constrangemos" e impelimos o modelo para uma atenção maior aos dados solicitados, tornando o modelo mais conservador diante de sequências não atestadas em seu repertório probabilístico. O protocolo estruturado não altera o acesso a conhecimento sociocultural externo, mas o critério interno de decisão sobre um repertório exclusivamente estatístico. A própria simulação da validação empírica constituiu, assim, uma manifestação da competência fraseológica artificial simulada.

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.

A dimensão cultural apresentou os resultados mais inferiores em todas as condições e em todas as categorias do corpus, e é essa uniformidade o fato que nos parece relevante. A média mais elevada registrada em qualquer configuração experimental foi de 1,45, obtida nas expressões opacas sob condição few-shot — valor que, em escala de 0 a 3, corresponde a pouco mais do que reconhecimento parcial da ancoragem sociocultural. Não há, portanto, célula em que o modelo tenha

demonstrado sensibilidade cultural satisfatória, mas há apenas graus de insuficiência. Ainda que o sistema reconheça referências culturais amplamente difundidas, como o Saci em "na pica do Saci", a explicitação de culturemas, de motivações históricas e de simbolismos permaneceu genérica ou ausente, o que aponta para a inexistência de um procedimento sistemático de identificação cultural e converge com as formulações de Dobrovol'skij e Piirainen (2005) e de Pamies Bertrán (2008) acerca do caráter compartilhado do conhecimento fraseológico pela comunidade.

Um dado que merece atenção é o da condição Chain-of-Thought, pois é a única em que as categorias de unidades convencionalizadas não estão sob efeito de teto, já que no few-shot todas atingiram desempenho máximo. Por isso, é possível observar que a taxa de análises validadas não se distribui de forma homogênea entre as categorias. Os somatismos apresentam o índice mais elevado do experimento, com 93,5% (29 de 31 itens), enquanto as unidades culturalmente marcadas registram o mais baixo, com 72,4% (21 de 29 Se a principal limitação do modelo estivesse ligada à ausência de corporalidade, seria razoável encontrar um desempenho inferior justamente nas unidades cuja motivação se ancora na experiência sensório-motora, isto é, nos somatismos. O que se observa, no entanto, é o contrário. Podemos concluir que esse padrão pode ser explicado na Teoria da Linguagem Figurativa Convencional: esquemas de imagem de base corporal tendem a apresentar alto grau de universalidade e transparência semântica e, como observa Pamies Bertrán (2008), os somatismos costumam ter significados mais claros, o que os torna recuperáveis a partir de regularidades distribucionais, sem necessidade de experiência corpórea direta. Já o simbolismo culturalmente sedimentado e as imagens complexas descritas por Dobrovol'skij e Piirainen (2005) dependem de conhecimento enciclopédico compartilhado por uma comunidade linguística específica, algo que não se reduz à coocorrência lexical. Portanto, a corporalidade continua sendo constitutiva da competência fraseológica humana, embora não seja ela que delimite, no desempenho da máquina, o que pode ou não ser recuperado. O fator decisivo é a

convencionalização sancionada por uma comunidade (dimensão que a arquitetura distribucional, por sua própria natureza, não consegue replicar).

Também destacamos a diferença entre as condições experimentais de prompts. Nas unidades convencionalizadas, a condição few-shot alcançou um desempenho potencializado, com as 105 expressões opacas, somáticas e culturais corretamente analisadas segundo a validação do pesquisador humano. Esse efeito, contudo, não se estende à detecção da não convencionalidade: nas contraprovas, a rejeição correta passou de 17,6% no zero-shot para 31,4% no few-shot, diferença não tão significativa. A superioridade do few-shot mostra-se, portanto, robusta quando o modelo pode apoiar-se em regularidades distribucionais já fossilizadas, mas ausente quando precisa decidir sobre a própria existência da unidade. A assimetria sugere que os exemplos fornecidos funcionam como reforço de padrões, e não como critério de validação.

16 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. Merece destaque o desempenho dos protocolos de Chain-of-Thought (CoT) nas contraprovas experimentais. Nessa condição, a taxa de rejeição correta alcançou 91,7% (22 de 24 itens com registro analítico recuperável; IC 95%: 73,0%–99,0%), contrastando com 31,4% no few-shot (11/35) e 17,6% no zero-shot (6/34). Entretanto, onze dos 35 itens da condição CoT não puderam ser recuperados, por erro de armazenamento, restringindo a cobertura da célula a 68,6%. Uma análise de sensibilidade por imputação extrema, considerando todos os itens ausentes como erros, reduziria a taxa para 62,9% (22/35), mas preserva a superioridade do CoT sobre o few-shot (teste exato de Fisher bilateral, $p = 0,016$). Já a diferença entre zero-shot e few-shot não é estatisticamente significativa ($p = 0,265$). Os resultados sugerem, portanto, que o ganho proporcionado pelos exemplos se

concentra na interpretação de unidades convencionalizadas, e não na detecção da não convencionalidade.

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.

17 Considerações finais

Este estudo, focado no Claude Haiku 4.5, com o objetivo de verificar se apresenta uma competência fraseológica artificial simulada comparável à humana, trouxe contribuições teóricas e metodológicas, além de evidenciar limitações.

No campo da Fraseologia, os resultados reforçam a importância da convencionalização como dimensão essencial da competência, já que recuperar apenas significados metafóricos não garante legitimidade. Para a Linguística Cognitiva, os dados mostram que o desempenho do modelo não se degrada de forma uniforme, mas seletiva, acompanhando não o grau de ancoragem sensório-motora, e sim a dependência de conhecimento culturalmente sedimentado. Assim, a corporalidade, embora constitutiva da competência humana, não é o fator decisivo; o papel central cabe à validação comunitária. Trata-se de uma contribuição empírica: no domínio fraseológico, os resultados favorecem o poder explicativo da Teoria da Linguagem Figurativa Convencional, sem invalidar a Teoria Cognitiva da Metáfora, que permanece fundamental para descrever os processos humanos.

Do ponto de vista metodológico, defendemos que avaliações baseadas apenas na acurácia semântica podem superestimar as capacidades dos modelos. É necessário incluir dimensões de convencionalização, adequação pragmática e sensibilidade cultural. Destacamos ainda o uso de contraprovas com expressões inexistentes e a validação por corpora e fontes lexicográficas externas. Os avanços observados após etapas de verificação sugerem que arquiteturas híbridas, combinando geração com consulta documental, podem reduzir significativamente as alucinações figurativas.

De modo aplicado, isso recomenda cautela no uso de LLMs sem supervisão humana em ensino, tradução, materiais didáticos e análise intercultural, pois explicações plausíveis podem mascarar falhas sistemáticas.

Entre as limitações deste presente estudo, ressaltamos: (i) foco em um único modelo (ii) apenas uma tentativa de aplicação por UF; (iii) recorte restrito ao português brasileiro. Futuras pesquisas poderão explorar outras línguas, comparar modelos abertos e fechados, investigar culturemas regionais, criar benchmarks

específicos e desenvolver protocolos que integrem avaliação qualitativa da adequação pragmática, sensibilidade sociocultural e coerência inferencial.

18 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.

19 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.

BENDER, Emily M.; GEBRU, Timnit; McMILLAN-MAJOR, Angelina; SHMITCHELL, Shmargaret**. On the dangers of stochastic parrots**: can language models be too big? In:

ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2021, Virtual Event. Proceedings [...]. New York: Association for Computing Machinery, 2021. p. 610-623. DOI: 10.1145/3442188.3445922.

BENDER, Emily M.; KOLLER, Alexander. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: **ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS**, 58., 2020. Proceedings [...]. Stroudsburg: Association for Computational Linguistics, 2020. p. 5185-5198. DOI: 10.18653/v1/2020.acl-main.463.

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.

MAHOWALD, Kyle; IVANOVA, Anna A.; BLANK, Idan A.; KANWISHER, Nancy; TENENBAUM, Joshua B.; FEDORENKO, Evelina. **Dissociating language and thought in large language models**. Trends in Cognitive Sciences, v. 28, n. 6, p. 517-540, 2024. DOI: 10.1016/j.tics.2024.01.011.

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.