Análise de similaridade semântica: Medida de desempenho entre modelos de IA Generativa da OpenAI e da DeepSeek

Gustavo Henrique MEDRADO[1]
Marcelo Marques ARAÚJO[2]
Pedro Franklin Cardoso SILVA[3]
Alexandre Gomes de SIQUEIRA[4]
Angela Maria GROSSI[5]

Universidade Estadual Paulista – UNESP, Universidade Federal de Uberlândia - UFU e Universidade da Flórida - UF

RESUMO
Este trabalho analisa a similaridade semântica aferida por dois modelos de Inteligência Artificial Generativa, o GPT-4o da OpenAI e o deepseek-v3, da DeepSeek, com o objetivo de medir a aptidão desses modelos no contexto da checagem de fatos. A partir de um conjunto de dados com 1.200 pares de entradas — compostos por checagens da agência Aos Fatos e respostas equivalentes geradas pelo GPT-4o, ambos os modelos classificaram as entradas em quatro faixas, de fraca à quase total, com base em duas técnicas de prompting: zero-shot e few-shot com cadeia de pensamento. Os resultados revelaram mais de 90% das respostas com medianas de similaridade entre boa (≥0.80) e quase total (0.95). Técnicas mais rigorosas de prompting, como o few-shot-CoT aplicado via DeepSeek, mostraram maior criticidade nas categorias de menor similaridade.

Palavras-chave: similaridade semântica; chatgpt; deepseek; inteligência artificial generativa; comunicação

A partir de um banco de dados brasileiro, medido paralelamente por modelos de IA Generativa estadunidense e chinês, este trabalho propõe uma classificação sintética do desempenho real das técnicas de prompting testadas, considerando tanto as métricas centrais (mediana de similaridade) quanto a capacidade crítica dos modelos frente às entradas mais ambíguas ou divergentes. Essa classificação integra técnicas de prompting (zero-shot e few-shot com cadeia de pensamento) e os modelos avaliadores (GPT-4o e DeepSeek-V3).

[1] Professor Substituto no Departamento de Letras, Linguística, Comunicação e Artes da UEMG Frutal. Doutorando em Mídia e Tecnologia pela UNESP. E-mail: gh.medrado@unesp.br
[2] Professor Adjunto da Faculdade de Educação - UFU, no curso de Jornalismo e no Programa de Pós-Graduação em Tecnologia, Comunicação e Educação. E-mail: marcelo.araujo@ufu.br
[3] Professor Adjunto do Instituto de Matemática e Estatística da UFU. E-mail: pedrofranklin@ufu.br
[4] Professor Assistente Instrucional do Departamento de Ciência da Computação e Informação e Departamento de Engenharia da Universidade da Flórida. E-mail: agomesdesiqueira@ufl.edu
[5] Professora Associada no Departamento de Jornalismo e no Programa de Pós-Graduação em Mídia e Tecnologia da UNESP Bauru. E-mail: angela.grossi@unesp.br

Metodologia

A abordagem metodológica adota um caráter experimental e empírico, amparado pelo método hipotético-dedutivo, e compreende a criação de um conjunto de dados original de 1.200 pares de entradas, extraídos de notícias verificadas pela agência Aos Fatos e de respostas geradas pelo modelo GPT-4o do ChatGPT. A análise dos dados envolveu a aplicação de medidas de similaridade semântica, métodos de clusterização (K-Means e Aglomerado Hierárquico) e testes estatísticos não-paramétricos, como o Teste do Sinal, para aferir o desempenho do modelo na verificação informacional.

Ao invés de realizar classificações binárias de afirmações como verdadeiras ou falsas, optou-se por uma abordagem baseada na análise da similaridade semântica entre os conteúdos gerados por um modelo de linguagem e os textos produzidos por uma agência de checagem. Essa escolha permite observar não apenas se o modelo acerta, mas como ele se aproxima, em conteúdo e intenção, de uma resposta qualificada.

Como o enfoque deste trabalho é analisar as potencialidades de LLMs, optou-se por utilizá-los também para essa medida de similaridade semântica. Numa escala de 0 a 1, foi solicitado que classificassem o quão próximo o conteúdo gerado pelo modelo está do conteúdo da agência de checagem.

Verificou-se como o modelo se comporta ao **comparar** informações jornalísticas. Essa comparação, realizada via API da OpenAI e da DeepSeek por meio do Google Colab na linguagem Python, permite entender como cada conteúdo desempenha e se são satisfatórios no quesito de checagem de informações. No Quadro 1 é possível ver um trecho dos códigos utilizados em cada um dos modelos avaliadores.

Quadro 1 - Exemplos otimizados dos prompts utilizados

Zero-Shot	***Few-Shot-CoT***
prompt = f""" Avalie a similaridade semântica entre os dois textos abaixo e retorne um número entre 0 e 1. O intuito é verificar se ambos os textos, ainda que escritos de maneiras diferentes, corroboram com a mesma ideia.	prompt = f""" Avalie a similaridade semântica entre os dois textos abaixo e retorne um número entre 0 e 1.Seu objetivo é identificar se os textos, mesmo escritos de maneiras diferentes, transmitem a mesma ideia. Escala de Similaridade: - 0.0 – 0.3: Fracos ou inexistentes paralelos semânticos - 0.31 – 0.6: Similaridade parcial - 0.61 – 0.85: Boa correspondência de sentido

- 1 significa que os textos têm o mesmo significado. - 0 significa que os textos não têm nenhuma relação semântica. - Valores intermediários representam diferentes graus de similaridade. Texto 1: "{text1}" Texto 2: "{text2}" Responda apenas com um número decimal entre 0 e 1, com duas casas decimais. Não inclua nenhuma outra informação na resposta. """	- 0.86 – 1.0: Correspondência quase total de conteúdo e intenção ### Exemplos: Texto 1: "[xxx]" Texto 2: "[xxx]" Analise foco, veracidade e fundamentação. Similaridade: **1.00** --- Texto 1: "[xxx]" Texto 2: "[xxx]" Analise foco, veracidade e fundamentação. Similaridade: **0.00** --- Agora avalie o seguinte par de textos: Texto 1: "{text1}" Texto 2: "{text2}" Pense passo a passo, identifique a similaridade e retorne apenas um número entre 0 e 1 com até duas casas decimais. **Não inclua nenhuma explicação ou justificativa.** """

Fonte: Elaboração própria para fins didáticos.

Todos os códigos, a redação dos prompts utilizados e o próprio *dataset* completo criado nesta pesquisa[6] estão disponibilizados nesta pasta do Google Drive[7].

Fundamentação teórica

De forma geral, prompt é uma informação enviada (entrada) a um modelo de IA Generativa que é utilizada como guia para gerar uma outra informação (saída), podendo se consistir em textos, imagens, sons ou outro conteúdo audiovisual (Meskó, 2023; White et al., 2023; Heston and Khun, 2023; Hadi et al., 2023; Brown et al., 2020 apud Schulhoff et al., 2024). Dentro de prompting, há as técnicas de prompting, que orientam como estruturar um ou uma sequência dinâmica de vários prompts. Dentre elas,

[6] Ao final, para a análise dessas comparações, o que inclui a clusterização e os testes de hipóteses, também foi utilizado a linguagem R, rodada no RStudio versão 2024.04.2, utilizando com base as bibliotecas readxl, dplyr e ggplot2.
[7] Códigos, Prompts e Dataset. Acesso em: https://drive.google.com/drive/folders/1HgCIE0VYp8aHTHFX0BuPuJ9JTrV-K_UF?usp=sharing

destacam-se duas, a few-shot e a zero-shot. A primeira faz parte da categoria intitulada ICL (In-Context Learning), ou aprendizagem no contexto, que refere-se à capacidade das IA Generativas em aprender habilidades e tarefas quando lhes são fornecidos exemplos e/ou instruções relevantes dentro do prompt (Schulhoff et al., 2024).

A segunda técnica, a zero-shot, não pertence à categoria ICL e não necessita de exemplos. Bastam simples direcionamentos, como o role-prompting, em que se consiste em atribuir uma função específica a IA Generativa.

Em ambas as técnicas, ainda é possível adicionar elementos que aprimoram o desempenho, como as cadeias de pensamento (Chain-of-Thought), ajudando os LLMs a articular seu raciocínio durante a solução de um problema (Zhang et al., 2023c apud Schulhoff et al., 2024). Isso inclui uma espécie de "passo a passo", guiando o LLM em como deve agir para realizar aquela tarefa.

Análise e/ou principais resultados e/ou contribuições da pesquisa

Para realizar a medição da similaridade semântica – expressão utilizada por essa pesquisa para parametrizar o quão os pares de texto, da agência e do LLM, são semelhantes e transmitem a mesma informação, embora redigidos de formas diferentes –, se valeu da aplicação dos dois tipos de *prompts* que tiveram o melhor desempenho no estudo de caso feito por Schulhoff et al. (2024), o *zero-shot* e o *few-shot CoT*.

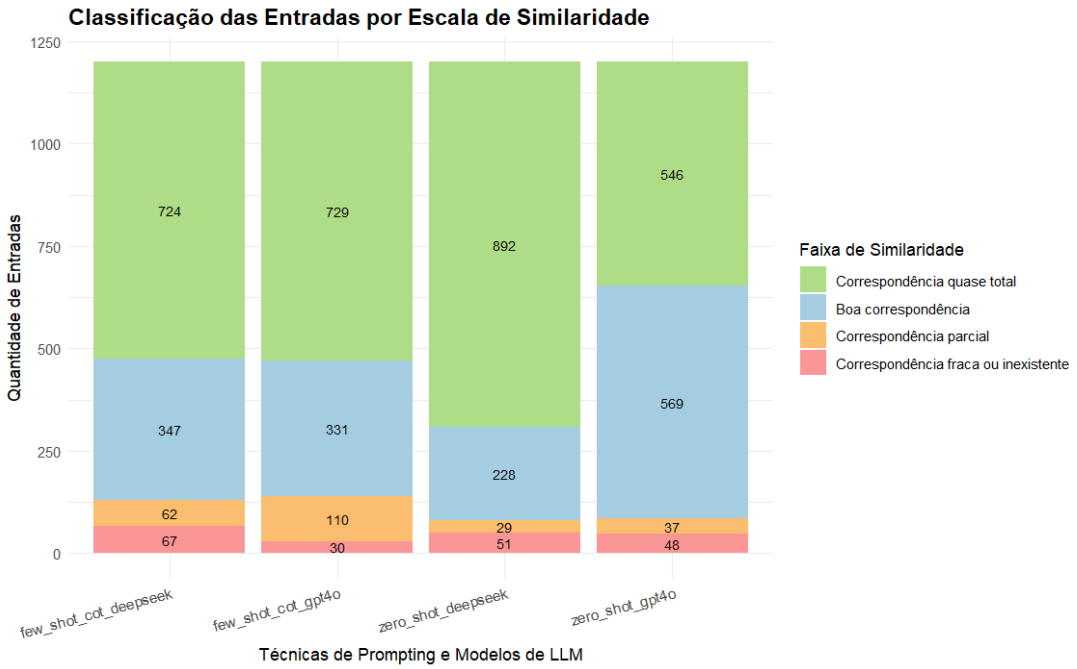
Em relação ao *zero-shot*, essa pesquisa aplicou o seu subtipo *Role Prompting*, que consiste em atribuir uma função específica a IA Generativa. Para esse trabalho, pediu que agisse como um " avaliador de similaridade semântica".

Já o *few-shot CoT*, recebeu no *prompting* dois exemplos de análise de similaridade, sendo um par com similaridade 1 e outro 0, além das instruções contendo a escala de similaridade a seguir. Também foi repassado a estrutura de pensamento passo a passo a seguir: **Foco:** Ambos os textos tratam da mesma alegação ou temática?; **Veracidade:** Ambos apresentam a mesma conclusão sobre a veracidade do conteúdo?; e **Fundamentação:** O raciocínio e as evidências utilizadas são semelhantes?

As técnicas de *prompting* validadas pelos testes de hipóteses classificaram as medianas da seguinte maneira: 0.95 para a few-shot-CoT e *zero-shot* da DeepSeek e, para as da OpenAI, 0.86 e 0.80 respectivamente. Mas esses resultados não podem ser

avaliados sozinhos. Para analisar o desempenho real de uma técnica de *prompting* é preciso considerar a capacidade delas em captar as nuances de cada entrada (Gráfico 1).

Gráfico 1 - Classificação dos Pares de Entradas por Escala de Similaridade



Fonte: Elaboração própria

Nas faixas **parcial** e **fraca ou inexistente**, os modelos, ao utilizar a técnica *zero-shot*, classificaram 40% menos pares de entradas do que ao utilizar a *few-shot CoT*. Ainda assim, essas técnicas possuem mediana igual ou superior às da *few-shot CoT*.

Os modelos *few-shot CoT* demonstraram ter maior criticidade e rigor nas avaliações, penalizando mais as entradas com maior divergência entre si. Embora essas observações ainda não sejam conclusivas e careçam de mais estudo, as evidências indicam que seja por causa das instruções mais detalhadas e exemplos fornecidos, característicos dessa técnica de *prompting*.

Conclusão

A few-shot-CoT com DeepSeek demonstrou correspondência quase total (mediana = 0.95) aliada a alta criticidade, especialmente na identificação de pares com

baixa convergência semântica, sendo a técnica mais equilibrada entre rigor e desempenho geral. A few-shot-CoT com GPT-4o alcançou correspondência quase total (mediana = 0.86), com comportamento semelhante ao anterior quanto ao rigor crítico, ainda que com desempenho ligeiramente inferior nas faixas superiores.

Já a técnica *zero-shot* com GPT-4o apresentou mediana, após teste de hipóteses, acima de 0.80, mas menor que a indicada inicialmente 0.86, o que colocou-a na faixa de boa correspondência. Seu desempenho teve maior estabilidade nas categorias altas, mas menor criticidade nas faixas de baixa similaridade, indicando uma tendência à suavização das avaliações. Por fim, o *zero-shot* com DeepSeek obteve mediana muito alta (0.95), mas com desequilíbrio evidente entre faixas e sensível ausência de rigor nas entradas mais problemáticas, o que sugere superestimação da similaridade semântica.

Para exemplificar melhor essa questão, recomenda-se verificar o trabalho completo (Medrado, 2026) com amostras randômicas de como cada técnica desempenhou ao medir a similaridade dos respectivos pares de entrada.

Essa síntese busca oferecer um panorama geral do comportamento das técnicas e modelos testados, apontando caminhos para escolhas mais assertivas no uso de LLMs voltados à verificação informacional. O desempenho não deve ser medido apenas pela média ou mediana, mas também pela capacidade de distinguir com precisão os casos de baixa similaridade, o que é essencial para aplicações críticas, como no Jornalismo de Prompt ou Prompt Jornalismo, campo no qual esta pesquisa está inserida.

REFERÊNCIAS

SCHULHOFF, Sander; ILIE, Michael; BALEPUR, Nishant; KAHADZE, Konstantine; LIU, Amanda; SI, Chenglei; LI, Yinheng; GUPTA, Aayush; HAN, HyoJung; SCHULHOFF, Sevien; DULEPET, Pranav Sandeep; VIDYADHARA, Saurav; KI, Dayeon; AGRAWAL, Sweta; PHAM, Chau; KROIZ, Gerson; LI, Feileen; TAO, Hudson; SRIVASTAVA, Ashay; DA COSTA, Hevander; GUPTA, Saloni; ROGERS, Megan L.; GONCEARENCO, Inna; SARLI, Giuseppe; GALYNKER, Igor; PESKOFF, Denis; CARPUAT, Marine; WHITE, Jules; ANADKAT, Shyamal; HOYLE, Alexander; RESNIK, Philip. **The prompt report: a systematic survey of prompting techniques**. arXiv preprint arXiv:2406.06608, 2024. https://doi.org/10.48550/arXiv.2406.06608

MEDRADO, Gustavo Henrique de Souza. **Inteligência Artificial e Informação Qualificada: Uma abordagem experimental para fundamentar o Jornalismo de Prompt**. 111 p. Dissertação (Mestrado em Tecnologia, Comunicação e Educação – Universidade Federal de Uberlândia. Uberlândia, 2026.