UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS–GRADUAÇÃO EM INFORMÁTICA APLICADA

DEIVISON GOMES DE AMORIM

MODELOS DE LINGUAGEM DE GRANDE ESCALA E ENGENHARIA DE PROMPT: UMA ABORDAGEM NA COMPARAÇÃO DE FRASES EM PORTUGUÊS BRASILEIRO.

RECIFE – PE

2024

DEIVISON GOMES DE AMORIM

MODELOS DE LINGUAGEM DE GRANDE ESCALA E ENGENHARIA DE PROMPT: UMA ABORDAGEM NA COMPARAÇÃO DE FRASES EM PORTUGUÊS BRASILEIRO.

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre em Informática Aplicada.

ORIENTADOR: Prof. Dr. Rafael Ferreira Leite de Mello

RECIFE – PE

2024

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Bibliotecária: Suely Manzi – CRB/4 - 809

A524m Amorim, Deivison Gomes de
Modelos de linguagem de grande escala e engenharia de PROMPT: uma abordagem na comparação de frases em português brasileiro / Deivison Gomes de Amorim. – 2024.
67 f.: il.

Orientador: Rafael Ferreira Leite de Mello.
Dissertação (Mestrado em Informática Aplicada) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2024.
Inclui bibliografia e apêndice(s).

1. Computação semântica 2. Processamento de linguagem natural (Computação) 3. Linguagem de programação (Computadores) – Semântica 4. Inteligência artificial I. Melo, Rafael Ferreira Leite de, orient. II. Título

CDD 004

Dedico este trabalho a minha mãe, Solange Gomes de Amorim. Ela, com muito esforço, criou meu irmão e eu e nos mostrou que a educação é o meio fundamental de transformação do ser e da sociedade.

Agradecimentos

Inicialmente, é de fundamental importância render graças e agradecimentos ao maior autor de qualquer obra científica: Deus! Sem Ele, em muitos momentos, a desistência e a descrença de finalizar seriam certas. O autor dos autores merece todos os agradecimentos aqui rendidos; sem a graça da inteligência e do discernimento, nada disso poderia ter sido realizado, muito menos concluído. Agradeço ao Amor por excelência e faço menção à frase de São João da Cruz: "No entardecer da vida, seremos julgados pelo amor."

Aproveito o ensejo para agradecer também à minha família. Sintetizo os agradecimentos centrados na pessoa da minha mãe, Solange Gomes de Amorim. Ela, com muita sabedoria e esforço, nos ajudou e ajuda na condução de nossas vidas, lutando arduamente para criar seus dois filhos, mesmo após a morte repentina de seu esposo, João Batista de Amorim, que menciono aqui "in memoriam". Mãe, a senhora merece o maior dos agradecimentos, pois sempre acreditou que, por meio da educação, poderíamos vencer na vida e nos tornarmos pessoas melhores. A senhora acertou em cheio ao pensar e orientar seus filhos nesse sentido. Estendo também esses agradecimentos à minha esposa, Mariana Sousa, e ao meu filho, José Miguel. Minha esposa sempre me incentivou na caminhada acadêmica e foi um apoio fundamental durante esse processo. Meu filho, José Miguel, a você, meus agradecimentos. Você foi o maior dos professores que pude ter, ensinando-me que amar é a atitude mais singela e nobre da vida humana. Obrigado por ser o maior sinal de amor em minha vida!

Agradeço aos amigos, concentrando todos na pessoa do Professor Mestre Rubens Karman de Paula Silva, meu amigo e irmão, Rubinho. Você foi e é um grande incentivador na vida acadêmica e sempre se alegrou comigo nas minhas conquistas. Com você, aprendi duas lições fundamentais para essa jornada: que é importante sempre iniciar e encerrar ciclos, e que, com dedicação e muito esforço, é possível alcançar os sonhos desejados. Obrigado por me incentivar e por ser um grande referencial em minha vida acadêmica.

Dedico também meus sinceros agradecimentos ao meu orientador, Professor Doutor Rafael Ferreira Leite de Mello. Professor, você foi muito mais que um orientador nesta caminhada; com você, aprendi, na prática, que educação e ciência se fazem acreditando nas pessoas, o que é uma atitude louvável. Em muitos momentos de dificuldade, você foi o

maior incentivador e acreditou até o fim que seria possível. O gesto mais nobre que posso expressar é o desejo de que meu filho tenha muitos professores dedicados e humanos como você em sua trajetória. Você é uma das melhores referências em minha vida acadêmica. Obrigado de coração, professor.

Aproveito também para agradecer ao Professor Doutor Ivaldir Honório de Farias Júnior. Tive a honra de conhecê-lo em 2016, durante uma especialização em Engenharia de Software. Na ocasião, foi meu orientador e seguiu me orientando em diversas nuances até os dias atuais. Professor, você é uma referência positiva para mim, tanto no meio acadêmico quanto na vida pessoal. Obrigado pelas orientações e por sempre se fazer presente.

Agradecer nunca é demais, portanto, agradeço imensamente às Irmãs Franciscanas de Nossa Senhora do Amparo, do Colégio Nossa Senhora do Amparo, em Surubim/PE. Direciono este agradecimento especial à Irmã Ercília Maria Bezerra Marinho. Irmã, a senhora foi uma das maiores incentivadoras na minha jornada no mestrado. Contribuiu, com gestos e ações, para que eu concluísse mais essa etapa da minha vida. Suas palavras de apoio e sua confiança no meu potencial foram combustíveis poderosos para a realização deste sonho. Obrigado por me ajudar a ser melhor; sua vocação fortalece a minha.

Agradeço aos meus alunos. Vocês me fazem compreender, na prática, que a educação é a base fundamental da sociedade. Ter a oportunidade de compartilhar um pouco do que sei com cada um de vocês é um grande motivador para continuar contribuindo com uma educação de qualidade, que alcance todos os setores da sociedade, sem distinção. Obrigado por estarem em meu caminho e por me lembrarem, a todo instante, que serei um eterno estudante também; vocês me ensinam muito.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

"Ou as pessoas e as empresas entendem Inteligência Artificial como uma nova dimensão da inteligência, não como um conjunto de ferramentas, de plataformas, de tecnologias, ou elas não vão sobreviver." .

(Silvio Meira)

Resumo

Este estudo investiga o impacto da Engenharia de Prompt na melhoria dos resultados obtidos por um Grande Modelo de Linguagem (LLMs), neste caso o ChatGPT, na tarefa de comparação de similaridade semântica textual entre frases em português brasileiro. A crescente popularização da Inteligência Artificial (IA) e o uso de Processamento de Linguagem Natural (PLN) têm gerado a necessidade de desenvolver técnicas que melhorem a interação entre humanos e máquinas, especialmente em termos de compreensão semântica. O trabalho foca na aplicação de engenharia de prompt, uma técnica que busca orientar a construção eficaz de comandos textuais (prompts) para que LLMs ofereçam respostas mais precisas e satisfatórias. Um uso inadequado de prompts pode limitar o potencial dos modelos de IA, resultando em respostas imprecisas. O objetivo principal da pesquisa é avaliar se o uso da engenharia de prompt melhora os resultados obtidos na comparação de similaridade semântica entre frases, utilizando um grande modelo de linguagem. Para isso, foram realizados experimentos envolvendo diferentes técnicas de engenharia de prompt e o ChatGPT, sendo os resultados avaliados com base em duas métricas reconhecidas: a Correlação de Pearson e o Erro Quadrático Médio (MSE). Esses resultados foram comparados com os obtidos no workshop ASSIN II, realizado em 2016, que focou na criação de modelos de IA para avaliar similaridade semântica e inferência textual em português. A pesquisa demonstra que a engenharia de prompt pode aumentar significativamente o desempenho de LLMs na tarefa de avaliação de similaridade semântica, contribuindo para o avanço do PLN e fomentando investigações científicas mais aprofundadas na área. Porém, para que essa melhoria ocorra de forma substancial, é necessário um processo sistemático de refinamento do prompt e testagem, assim obtém-se um melhor retorno por parte do modelo. Ao otimizar o uso dessas ferramentas, este estudo não apenas promove o desenvolvimento de técnicas mais eficazes, mas também fortalece a aplicação da IA na sociedade e na ciência.

Palavras-chave: Similaridade Semântica Textual. Processamento de Linguagem Natural. Grandes Modelos de Linguagem. Engenharia de Prompt

Abstract

This study investigates the impact of Prompt Engineering on improving the results obtained by a Large Language Model (LLM), in this case ChatGPT, in the task of comparing textual semantic similarity between sentences in Brazilian Portuguese. The growing popularization of Artificial Intelligence (AI) and the use of Natural Language Processing (NLP) has generated the need to develop techniques that enhance human-machine interaction, especially in terms of semantic understanding. The work focuses on the application of prompt engineering, a technique aimed at guiding the effective construction of textual commands (prompts) so that LLMs provide more accurate and satisfactory responses. Inadequate use of prompts can limit the potential of AI models, resulting in inaccurate responses. The main objective of the research is to assess whether the use of prompt engineering improves the results obtained in the comparison of semantic similarity between sentences, using a large language model. To achieve this, experiments were conducted involving different prompt engineering techniques and ChatGPT, with the results evaluated based on two recognized metrics: Pearson Correlation and Mean Squared Error (MSE). These results were compared to those obtained in the ASSIN II workshop, held in 2016, which focused on the creation of AI models to assess semantic similarity and textual inference in Portuguese. The research demonstrates that prompt engineering can significantly enhance the performance of LLMs in the task of evaluating semantic similarity, contributing to the advancement of NLP and fostering deeper scientific investigations in the field. However, for this improvement to occur substantially, a systematic process of prompt refinement and testing is necessary to achieve better responses from the model. By optimizing the use of these tools, this study not only promotes the development of more effective techniques but also strengthens the application of AI in society and science.

Keywords: Textual Semantic Similarity. Natural Language Processing. Large Language Models. Prompt Engineering

Lista de Figuras

Figura 1 – Matriz de Correlação 44
Figura 2 – Resultados Erro Quadrático Médio 46
Figura 3 – Resultados em Correlação de Pearson 47
Figura 4 – Resultados em Erro Quadrático Médio - MSE 48
Figura 5 – Resultados Anotados X Resultados Preditos - Prompt Nível 01 51
Figura 6 – Resultados Anotados X Resultados Preditos - Prompt Nível 05 52

Lista de tabelas

Tabela 1 – Etapas da Metodologia 27
Tabela 2 – Resultados Prompts Nível 01 37
Tabela 3 – Resultados Prompts Nível 02 37
Tabela 4 – Resultados Prompts Nível 03 38
Tabela 5 – Resultados Prompts Nível 04 38
Tabela 6 – Resultados Prompts Nível 05 39
Tabela 7 – Resultados com Maior Proximidade dos Valores Anotados 40
Tabela 8 – Resultados Iguais aos Valores Anotados 40
Tabela 9 – Resultado - Correlação de Pearson. 43
Tabela 10 – Resultado - Erro Quadrático Médio (MSE). 45

Lista de Siglas

IA	Inteligência Artificial
PLN	Processamento de Linguagem Natural
MSE	Mean Squared Error
LLM	Large Language Model

Sumário

1 Introdução 13
1.1 Pergunta de Pesquisa 15
1.2 Objetivos 15
1.2.1 Objetivo Geral 15
1.2.2 Objetivos Específicos 15
1.3 Organização do Trabalho 16
2 Fundamentação Teórica 18
2.1 Inteligência artificial 18
2.2 Processamento de linguagem natural (PLN) 19
2.3 Inteligência Artificial Generativa 21
2.4 ChatGPT 22
2.5 Engenharia de prompt 23
3 Trabalhos Relacionados 25
4 Metodologia 27
4.1 Pesquisa Bibliográfica 28
4.2 Desenvolvimento Experimental 29
4.3 Base de Dados 29
5 Resultados e Discussões 31
5.1 Construção e Análise de Prompts 31
5.2 Desempenho dos Prompts 36
5.3 Resultados da Correlação de Pearson 41
5.4 Erro Quadrático Médio (MSE) 45
5.5 Resultados da Pesquisa X Resultados do ASSIN II 46
5.6 Discussões 49
5.6.1 Resultados Anotados X Resultados Preditos 51
6 Considerações Finais 54
6.1 Limitações da pesquisa 55
6.2 Trabalhos Futuros 56
Referências 57
7 Apêndices 60

7.0.1 Evolução dos Prompts 60
7.0.2 Prompts Utilizados nos Experimentos 62

1 Introdução

O uso de técnicas de inteligência artificial (IA) tem se tornado cada vez mais frequente nos dias atuais. Diversas áreas já apresentam perspectivas que podem ser significativamente beneficiadas por ferramentas que utilizam IA (TAVARES et al., 2020). Esse impacto promove a necessidade de um número crescente de estudos que explorem essa área, de modo que a sociedade possa usufruir de softwares inteligentes de forma mais ampla. O surgimento de abordagens que facilitam o entendimento e o uso dessas ferramentas contribui para popularizar o acesso à IA. Um exemplo notável é o ChatGPT, que permite que seus usuários acessem seu poder computacional por meio de textos escritos em formato de chat, promovendo uma interação simples e acessível.

Nesse contexto, surge a necessidade de aprimorar essas ferramentas, principalmente no que diz respeito à capacidade de entender a linguagem humana. A área de IA conhecida como Processamento de Linguagem Natural (PLN) se destaca nesse cenário, uma vez que estuda como ocorre a interação entre humanos e máquinas a partir da linguagem. Dentre as várias subáreas do PLN, merece destaque a vertente de avaliação semântica textual, que investiga como os softwares de inteligência artificial compreendem o sentido do que é escrito, um desafio significativo, considerando a ampla gama de possibilidades que um texto pode gerar semanticamente.

Com a crescente disseminação da IA de forma acessível, também surge a necessidade de desenvolver formas mais eficazes para explorar todo o potencial dessas ferramentas inteligentes. Softwares como o ChatGPT são classificados como Grandes Modelos de Linguagem, ou LLMs (Large Language Models, em inglês). Esses modelos são caracterizados pela interação direta entre o usuário e a ferramenta, sendo, portanto, crucial o uso eficaz da forma de ”diálogo”entre ambos. Nesse sentido, surge a área da Engenharia de Prompt, que se dedica a orientar os usuários sobre as melhores formas de construir os textos de interação com os LLMs. Esses textos são conhecidos como “prompts”. O uso inadequado ou incoerente de prompts pode resultar em respostas imprecisas ou insatisfatórias por parte do modelo, criando uma falsa impressão de que a necessidade foi atendida. Além disso, a construção ineficaz dos prompts limita o potencial do modelo, que acaba utilizando apenas uma pequena parte de seu poder de processamento devido à simplicidade do comando inserido.

Dessa forma, o presente estudo tem como objetivo avaliar se o uso da engenharia de prompt melhora os resultados obtidos na comparação de similaridade semântica entre frases, utilizando um grande modelo de linguagem. Para alcançar esse objetivo, será fundamental testar diferentes técnicas de engenharia de prompt, buscando identificar os melhores resultados gerados a partir de sua utilização. Além disso, a pesquisa utilizará grandes modelos de linguagem (LLM) na tarefa de comparação de similaridade semântica textual entre frases, e buscará relacionar o uso da engenharia de prompt à possível melhoria dos resultados obtidos. Assim, a pesquisa se propõe a responder à seguinte pergunta: "A utilização da engenharia de prompt em modelos de linguagem de grande escala (LLM) melhora a eficiência dos resultados na comparação entre frases em português brasileiro?"

Evidencia-se, portanto, que estudos no campo do Processamento de Linguagem Natural (PLN) são essenciais para o avanço dessa área. Além de contribuírem para o desenvolvimento de ferramentas e técnicas mais robustas, esses estudos fortalecem a comunidade científica ao fomentar investigações mais aprofundadas. No contexto brasileiro, eventos como o workshop de Avaliação de Similaridade Semântica e de Inferência Textual, conhecido como ASSIN II, têm desempenhado um papel importante no avanço dessa área. Realizado em 2016, o ASSIN II teve como objetivo criar modelos de inteligência artificial que apresentassem melhor performance na avaliação de similaridade semântica e inferência textual, abordando, especificamente, as variantes da língua portuguesa do Brasil e de Portugal.

Para a realização desta pesquisa, foi inicialmente conduzida uma pesquisa na literatura sobre o tema. A partir dessa base teórica, foram realizados experimentos práticos envolvendo a elaboração de diferentes prompts e a testagem com um LLM, especificamente o ChatGPT. Os resultados dos experimentos foram medidos utilizando duas métricas amplamente reconhecidas nos estudos de modelos de inteligência artificial: a Correlação de Pearson e o Erro Quadrático Médio (MSE). Com os resultados em mãos, foi realizada uma comparação entre as respostas obtidas nesta pesquisa e os resultados do ASSIN II, a fim de buscar evidências adicionais sobre a eficácia da Engenharia de Prompt na melhoria dos resultados de Grandes Modelos de Linguagem na tarefa de comparação de similaridade semântica textual.

Em síntese, este estudo reforça a importância da engenharia de prompt como

um fator que pode potencializar significativamente o desempenho dos grandes modelos de linguagem, ao passo que contribui para o avanço do Processamento de Linguagem Natural e suas aplicações na sociedade e na ciência.

1.1 Pergunta de Pesquisa

A utilização de engenharia de prompt em modelos de linguagem de grande escala (LLM) melhora a eficiência dos resultados na comparação entre frases em português brasileiro?

1.2 Objetivos

Nesta seção, serão demostrados os objetivos norteadores dessa presente pesquisa. Abaixo destaca-se o objetivo geral que norteará todo o enfoque do trabalho e os objetivos específicos que serão etapas fundamentais na construção do conhecimento aqui estudado e apresentado.

1.2.1 Objetivo Geral

- Avaliar se o uso de engenharia de prompt, em grandes modelos de linguagem (LLM), melhora os resultados obtidos acerca da comparação de similaridade textual entre frases em língua portuguesa brasileira.

1.2.2 Objetivos Específicos

- Testar diferentes técnicas de engenharia de prompt na busca pelo melhor resultado obtido a partir da sua utilização;
- Utilizar o Grande Modelo de Linguagem (LLM), ChatGPT, na atividade de comparação de similaridade semântica textual entre frases.
- Relacionar o uso de engenharia de prompts à possível melhoria dos resultados gerados por um Grande Modelo de Linguagem na tarefa de comparação semântica textual.

1.3 Organização do Trabalho

Além deste capítulo inicial, que tem como objetivo apresentar uma introdução geral ao tema da pesquisa, este trabalho está estruturado em mais seis capítulos, organizados da seguinte maneira:

O capítulo 2 aborda os conceitos gerais relacionados à temática investigada. Nele, são discutidos os fundamentos teóricos essenciais que sustentam e direcionam os temas abordados ao longo do trabalho, oferecendo ao leitor uma base sólida de conhecimento para a compreensão das questões discutidas. Este capítulo tem um papel fundamental, pois estabelece as bases conceituais que dão suporte às discussões subsequentes.

O capítulo 3 é dedicado à explanação dos trabalhos relacionados à pesquisa em questão. Nesse capítulo, são apresentados os estudos e as construções científicas que serviram de alicerce para a realização deste estudo. São analisados trabalhos anteriores que influenciaram e nortearam o desenvolvimento da pesquisa, possibilitando uma melhor compreensão do estado da arte e das lacunas existentes no campo, além de justificar as escolhas metodológicas e teóricas feitas pelo autores.

O capítulo 4 descreve detalhadamente as metodologias adotadas para a execução da pesquisa. Neste ponto, o trabalho se aprofunda na explicação das etapas seguidas para a realização dos experimentos e das análises. Cada procedimento foi cuidadosamente delineado para assegurar a precisão dos resultados e a construção sólida do conhecimento científico que esta dissertação se propôs a gerar e discutir. Aqui, são abordadas as técnicas de coleta de dados, o processo de experimentação e a estratégia de análise aplicada ao longo do estudo.

No capítulo 5, são apresentados os resultados obtidos com os experimentos e estudos realizados, acompanhados de uma análise crítica e reflexiva sobre os dados. Este capítulo foca nas discussões e interpretações dos resultados, confrontando-os com os objetivos estabelecidos anteriormente, e apontando os principais achados da pesquisa. São também exploradas as implicações dos resultados para a área de estudo e para o campo científico como um todo.

Já o capítulo 6 oferece as considerações finais deste trabalho. Nele, são apresentadas as conclusões derivadas da pesquisa, destacando o que foi possível observar e aprender ao longo do estudo. Além disso, são discutidas as limitações do trabalho, o que abre espaço para sugestões de pesquisas futuras, sinalizando as possibilidades de

continuidade dos estudos e a ampliação das discussões iniciadas neste trabalho.

Finalmente, o capítulo 7 é dedicado aos apêndices. Este capítulo reúne documentos e materiais complementares que são fundamentais para a compreensão completa da pesquisa, mas que, por sua natureza detalhada ou técnica, foram alocados separadamente. Esses apêndices oferecem suporte adicional aos argumentos e às evidências apresentadas nos capítulos anteriores, enriquecendo ainda mais o trabalho.

2 Fundamentação Teórica

2.1 Inteligência artificial

Uma das grandes inovações que surgiram com o crescente avanço tecnológico é a capacidade de agentes computacionais pensarem de forma semelhante aos seres humanos (PEREIRA et al., 2023). Esse campo é denominado inteligência artificial (IA), que combina um imenso poder de processamento computacional com a capacidade de compreender e analisar vastas bases de dados provenientes das mais diversas áreas. Segundo (SOUZA, 2022), o avanço tecnológico que vivemos atualmente tem o potencial de revolucionar a forma como adquirimos e utilizamos a informação no dia a dia. Diversas áreas do conhecimento já se beneficiam do uso da IA, aplicando-a em processos que vão desde a detecção de doenças a partir de exames médicos até estratégias para aumentar as vendas de um segmento comercial específico, utilizando dados coletados de experiências anteriores.

A inteligência artificial assume um papel proeminente nos dias atuais, como destacado por (LUDERMIR, 2021). A autora enfatiza que as máquinas não apenas desempenham tarefas manuais, mas também atividades que exigem o uso do que consideramos inteligência humana. Esse campo se dedica à criação de rotinas computacionais que não apenas auxiliam, mas frequentemente substituem o esforço humano, inclusive em tarefas complexas, exaustivas e que demandam grande precisão. A aplicação da IA tornou-se abrangente, estando presente em diversos segmentos. Um exemplo notável é a operação segura de veículos por software, minimizando riscos à vida humana, conforme afirmado por (FENG et al., 2021).

No entanto, é imprescindível destacar que o uso da inteligência artificial requer muito cuidado, disciplina e bom senso (FERRARO; COELHO, 2024). O estudo sobre as capacidades de uma IA está em constante expansão, o que implica a necessidade de cautela na forma como esses algoritmos inteligentes atuam e impactam a sociedade. (FENG et al., 2021) ressaltam que, para que um software que controla um automóvel resguarde a vida humana de forma eficaz, é necessário um treinamento adequado, utilizando dados corretos e prevenindo falhas. A responsabilidade e a ética no desenvolvimento e aplicação de IA são fundamentais para garantir a segurança e o bem-estar das pessoas.

Outro campo em que o uso da IA é proeminente é na interpretação da linguagem humana. Esta tarefa é de alta complexidade, pois exige que o algoritmo perceba as nuances da comunicação humana. O processamento de linguagem natural (PLN), conforme explicado por (OTTER et al., 2020), é a área da IA responsável pela interpretação e manipulação textual. O avanço da inteligência computacional nessa área traz amplas possibilidades de melhoria para processos que antes eram realizados exclusivamente por humanos. O PLN permite até que algoritmos avaliem redações humanas para verificar se estão alinhadas com o tema proposto, tudo isso de forma automatizada e com mínima interferência humana, de forma bastante eficaz (PINHO et al., 2022).

A evolução tecnológica impulsionada pela inteligência artificial continua a moldar nosso cotidiano. Desde a execução de tarefas práticas até a interpretação da linguagem humana, essa área continua a expandir suas fronteiras, explorando novas possibilidades e desafios para a interação entre humanos e máquinas. A tendência é que a IA se torne cada vez mais integrada aos nossos processos diários, trazendo melhorias significativas e transformando a maneira como vivemos e trabalhamos. A contínua pesquisa e desenvolvimento nesse campo são essenciais para garantir que essas tecnologias sejam utilizadas de forma ética e eficaz, promovendo benefícios para toda a sociedade (SARTO et al., 2024).

2.2 Processamento de linguagem natural (PLN)

O Processamento de Linguagem Natural (PLN) é uma das ramificações mais significativas dentro do vasto campo da inteligência artificial (IA). Essa área, como destacado por (NASCIMENTO, 2024), é essencialmente definida como um conjunto de técnicas destinadas a compreender tanto a estrutura quanto o significado presentes em textos, com o objetivo de reconhecer e gerar linguagem natural. O foco principal do PLN reside na compreensão semântica de textos por parte de modelos de IA, o que possibilita a geração de respostas que se aproximam cada vez mais da linguagem e da lógica humana, elevando a usabilidade dos sistemas que operam com base nesses algoritmos.

Dentro das inúmeras capacidades proporcionadas pelo PLN, destaca-se a avaliação da similaridade semântica textual. De acordo com (FONSECA et al., 2016b), a similaridade semântica textual pode ser entendida como ”uma medida numérica,

geralmente variando de 1 a 5, que expressa o grau de similaridade entre o conteúdo de duas sentenças". Porém, a definição exata desta tarefa não é universal e outros tipos de corpus podem carregar medidas diferentes. Esta avaliação é fundamental, pois vai além da simples correspondência de palavras, englobando a interpretação do contexto e da intenção subjacente às expressões utilizadas.

Os estudos sobre similaridade semântica textual desempenham um papel crucial em diversas aplicações práticas, que abrangem desde a detecção de plágio até a recuperação de informação relevante em grandes bases de dados (PINHO et al., 2022). Por exemplo, na detecção de plágio, a similaridade semântica permite identificar casos em que o conteúdo foi reescrito, mas o significado essencial permanece inalterado. Na recuperação de informação, essa métrica aprimora a relevância dos resultados de busca, garantindo que as respostas estejam alinhadas com a intenção do usuário, mesmo quando a consulta não coincide exatamente com o conteúdo dos documentos.

Além disso, a similaridade semântica textual é vital para o desenvolvimento de sistemas de resposta a perguntas, nos quais a precisão da correspondência semântica entre a pergunta e as respostas disponíveis é crucial para fornecer respostas corretas e contextualmente adequadas (TORRES et al., 2023). Na análise de sentimentos, essa avaliação permite que os algoritmos compreendam melhor as nuances emocionais expressas em textos, distinguindo entre sentimentos similares expressos de maneiras diferentes. Ademais, na tradução automática, a similaridade semântica é utilizada para garantir que a tradução preserve o significado original, mesmo quando as frases são estruturadas de forma diferente no idioma de destino.

Em suma, os estudos e as aplicações da similaridade semântica textual não apenas aprimoram a capacidade dos sistemas de IA em entender e processar a linguagem humana, mas também impulsionam a qualidade da interação entre humanos e máquinas, tornando essa comunicação cada vez mais natural e eficaz. Dessa forma, o PLN não só possibilita avanços técnicos, mas também promove uma integração mais fluida e eficiente das tecnologias de IA no cotidiano das pessoas.

2.3 Inteligência Artificial Generativa

O avanço tecnológico na área de inteligência artificial trouxe consigo a criação dos Grandes Modelos de Linguagem, conhecidos como LLMs (Large Language Models). Esses modelos são caracterizados por terem sido treinados em um vasto volume de dados textuais, o que lhes confere uma notável capacidade de gerar textos coesos e contextualizados. Segundo (NAVEED et al., 2023), LLMs são ”sistemas inteligentes baseados em inteligência artificial que conseguem processar e produzir respostas textuais com uma comunicação coerente, além de generalizar diferentes tarefas”.

O funcionamento desses modelos é fundamentado na previsão da palavra seguinte em uma frase, com base nas palavras anteriores. Esse processo de predição contínua permite que o modelo gere textos altamente contextuais, mantendo a coesão e a coerência ao longo do conteúdo produzido. Essa habilidade de prever e gerar texto coerente torna os LLMs extremamente versáteis, podendo ser aplicados em diversas áreas, desde a criação de conteúdos com padrões semelhantes à escrita humana até a produção de textos técnicos e de alto nível.

A introdução de modelos tão complexos e poderosos transformou radicalmente o uso de inteligência artificial, tornando-os acessíveis e fáceis de utilizar. Na atualidade, muitos desses modelos são integrados a interfaces amigáveis, projetadas para facilitar a interação e compreensão, permitindo que um número significativamente maior de usuários, independentemente de seu nível de instrução, possa usufruir de seus benefícios (NETO; MIERS, 2024). Essa democratização do acesso aos LLMs expandiu suas aplicações para um público mais amplo, oferecendo possibilidades antes restritas a especialistas.

Entre as várias capacidades oferecidas pelos Grandes Modelos de Linguagem, destaca-se a habilidade de transformar entradas simples em respostas elaboradas e complexas. O usuário pode inserir comandos textuais básicos, e o modelo, por meio de processamento avançado, responde com retornos que simulam ou replicam a qualidade de uma resposta humana. Um exemplo prático dessa capacidade é a verificação de similaridade semântica textual, uma ferramenta crucial para a detecção de plágio em textos. Ao identificar semelhanças entre diferentes textos, os LLMs ajudam a combater essa prática, que pode configurar não apenas uma violação ética, mas também um crime, dependendo do contexto e da legislação aplicável.

Além disso, a crescente sofisticação desses modelos amplia suas possibilidades de

uso em outras áreas do Processamento de Linguagem Natural (PLN), onde a precisão e a capacidade de gerar respostas contextualizadas são essenciais. Esses avanços têm o potencial de revolucionar a maneira como a inteligência artificial é utilizada para a interpretação e produção de linguagem natural, contribuindo para uma interação mais eficaz entre humanos e máquinas.

2.4 ChatGPT

O ChatGPT, desenvolvido pela OpenAI, é uma ferramenta avançada de inteligência artificial projetada para gerar respostas de alto nível a partir da interação com humanos. Utilizando um vasto conjunto de dados e uma interface de chat acessível, o sistema possibilita que seus usuários insiram comandos de diferentes níveis de complexidade, recebendo em retorno textos coerentes e contextualizados (SANT´ANA et al., 2023). De acordo com (FILHO et al., 2023), "O ChatGPT destaca-se pela sua capacidade de compreender linguagem natural e gerar respostas a partir de perguntas feitas pelo usuário".

Com seu surgimento, o ChatGPT provocou uma série de debates sobre seu impacto na sociedade. Uma questão que emergiu foi o potencial de plágio nos textos gerados, dado que ainda há incerteza sobre a originalidade das respostas criadas pelo modelo e se elas podem ser classificadas, de alguma maneira, como conteúdo plagiado. Além disso, a qualidade das respostas geradas também está em discussão. (LO, 2023) aponta que "o modelo, no estado atual, pode gerar informações incorretas ou contraditórias, embora estas aparentem estar corretas".

Apesar desses desafios, o ChatGPT tem se mostrado um diferencial relevante, especialmente em aplicações específicas que aproveitam suas capacidades. No campo educacional, por exemplo, a ferramenta tem facilitado o acesso ao conhecimento, sendo uma aliada tanto de professores, que podem utilizá-la para preparar aulas, quanto de estudantes, que encontram no sistema uma fonte de apoio para o aprendizado e aprofundamento em diversas disciplinas (BARBOSA et al., 2024).

Neste estudo, o sistema da OpenAI foi empregado com o objetivo de verificar a similaridade semântica textual entre pares de sentenças, demonstrando um alto potencial para essa tarefa específica. Os resultados obtidos indicam uma performance superior, em alguns casos, em comparação a outros modelos utilizados para o mesmo fim, evidenciando

a eficácia do ChatGPT na análise e processamento de linguagem natural.

Com a adoção direcionada, as capacidades dessa ferramenta tornam-se cada vez mais indispensáveis, ampliando suas possibilidades de uso em diversas áreas do conhecimento e abrindo novos caminhos para a interação entre humanos e máquinas.

2.5 Engenharia de prompt

O uso de assistentes conversacionais, especialmente aqueles desenvolvidos com base em Grandes Modelos de Linguagem (LLMs), como o ChatGPT, tem demonstrado um impacto significativo na forma como a tecnologia da informação é aplicada em diversos campos do conhecimento. Esses assistentes, em contraste com os chatbots tradicionais, são capazes de superar as limitações de interações previsíveis e respostas pré-programadas. Enquanto os chatbots convencionais frequentemente geram insatisfação nos usuários devido à rigidez de suas respostas, os assistentes baseados em LLMs destacam-se pela flexibilidade, adaptação e a capacidade de gerar respostas em linguagem natural. Isso cria uma experiência de interação mais humanizada e satisfatória, favorecendo um engajamento mais profundo dos usuários com a tecnologia, conforme ressaltado por (SHARMA et al., 2022).

A engenharia de prompts surge como uma área essencial no desenvolvimento dessas interações sofisticadas com os modelos de linguagem. O principal objetivo dessa área é desenvolver formas eficazes de comunicação entre o usuário e os LLMs, criando prompts que direcionam o comportamento do modelo de forma precisa e eficiente. Um prompt, nesse contexto, pode ser descrito como uma instrução textual cuidadosamente elaborada para orientar o modelo na geração de respostas que sejam relevantes e contextualmente apropriadas, pensamento que ratifica o exposto por (SILVA et al., 2024) que afirma que “essas estratégias consistem em fornecer um exemplo com a sequência de passos de raciocínio para auxiliar na resolução da tarefa solicitada fazendo com que o modelo aprenda não apenas a gerar respostas, mas a elaborar uma sequência coerente de passos [...]”. Sendo assim, a engenharia de prompts visa maximizar a performance do modelo, extraindo respostas mais acuradas e alinhadas com as necessidades do usuário.

Segundo (NASCIMENTO, 2024), a engenharia de prompts abrange diversas técnicas e abordagens, cada uma com suas particularidades e aplicabilidades específicas.

Nesta pesquisa, duas dessas técnicas foram utilizadas para explorar a eficácia dos prompts em modelos LLMs. A primeira delas, a técnica de Zero-Shot, consiste em criar prompts nos quais o modelo é solicitado a realizar tarefas sem exemplos prévios. Isso significa que o modelo precisa confiar exclusivamente em seu vasto conhecimento geral, adquirido durante o treinamento, para fornecer uma resposta adequada. Esse método, embora desafiador para o modelo, testa sua capacidade de generalização. A segunda técnica aplicada foi a Few-Shot, onde o modelo recebe alguns exemplos para orientar a tarefa. Nesse caso, os exemplos fornecidos atuam como guias, permitindo que o modelo compreenda melhor o padrão desejado para a tarefa e aumente a precisão das respostas geradas. A técnica Few-Shot visa aprimorar a eficiência do modelo, oferecendo-lhe referências claras de como responder adequadamente às demandas do usuário. Ambas as técnicas, Zero-Shot e Few-Shot, desempenham papéis cruciais na maximização do potencial dos LLMs, permitindo um desempenho adaptativo e refinado em diferentes situações e contextos de interação.

3 Trabalhos Relacionados

Este capítulo faz menção aos trabalhos científicos que embasaram e deram subsídio para o desenvolvimento da presente dissertação. Destacam-se três trabalhos que serviram de pilar na organização dos trabalhos realizados, eles foram primordiais para que fosse atingidos os objetivos estabelecidos.

O campo de estudo envolvendo Grandes Modelos de Linguagem (LLMs) tem se expandido significativamente nos últimos anos, com avanços notáveis em áreas como raciocínio lógico e tarefas de linguagem natural complexas. Uma importante contribuição foi apresentada por (KOJIMA et al., 2022), onde o foco está em uma técnica chamada Chain of Thought (CoT). Esta técnica, baseada na criação de prompts que induzem raciocínios passo a passo, demonstrou resultados impressionantes em tarefas de aritmética e raciocínio simbólico. O estudo propõe que os LLMs, frequentemente vistos como few-shot learners, possuem também capacidades de raciocínio zero-shot que podem ser exploradas por meio de prompts adequados. Ao adicionar a frase ”Vamos pensar passo a passo”antes das respostas, os modelos testados obtiveram um desempenho significativamente melhor em benchmarks complexos, sem a necessidade de exemplos previamente ajustados. Essa descoberta ressalta o potencial inexplorado dos LLMs, que podem ser melhor compreendidos e aproveitados com abordagens mais simples para a elaboração de prompts.

Outro trabalho relevante, de (SANTU; FENG, 2023), aborda os desafios de avaliar o desempenho dos LLMs em tarefas complexas. A pesquisa destaca a grande variação de performance dos modelos dependendo do tipo de prompt utilizado, o que torna difícil a padronização das avaliações. Para enfrentar esse problema, o autor propõe uma taxonomia geral para a criação de prompts, com o objetivo de facilitar estudos comparativos e aprimorar a análise do desempenho dos LLMs. A padronização sugerida pelo estudo permite uma melhor comparação entre diferentes abordagens e modelos, contribuindo para o desenvolvimento de benchmarks mais consistentes e precisos para tarefas complexas de processamento de linguagem natural.

Além desses avanços, (FONSECA et al., 2016a) investigou a aplicação de modelos de inteligência artificial na tarefa de inferência textual e de similaridade semântica, utilizando dados da avaliação conjunta ASSIN. O estudo se concentra na língua portuguesa

e apresenta um corpus anotado para essas tarefas. A pesquisa traz inovações ao incluir três classes distintas na tarefa de inferência textual (Implicação, Paráfrase e Nenhuma das duas), algo incomum em outros estudos da área. Diferentes estratégias foram exploradas pelas equipes participantes, resultando em uma análise abrangente das abordagens possíveis para essas tarefas no contexto do português.

Em relação ao presente trabalho, que trata da comparação de frases em português brasileiro utilizando técnicas de engenharia de prompt e modelos de linguagem de grande escala, os estudos revisados fornecem uma base sólida. A pesquisa aqui apresentada se beneficia das técnicas de elaboração de prompts exploradas em (KOJIMA et al., 2022), especialmente no que diz respeito à utilização de prompts de raciocínio passo a passo para melhorar o desempenho dos LLMs em tarefas complexas. A padronização sugerida por (SANTU; FENG, 2023) também é relevante, pois pode auxiliar na comparação eficaz de diferentes estratégias de prompts em língua portuguesa do Brasil, garantindo uma avaliação mais precisa. Por fim, o foco de (FONSECA et al., 2016a) na inferência textual e na similaridade semântica fornece uma base metodológica que será aplicada na análise comparativa entre frases, alinhada aos objetivos deste estudo no contexto da língua portuguesa.

4 Metodologia

Tabela 1 – Etapas da Metodologia

Etapa	Atividade Desenvolvida
1ª	Pesquisa Bibliográfica
2ª	Desenvolvimento Experimental
3ª	Utilização da Base de Dados
4ª	Construção e Análise de Prompts
5ª	Desempenho dos Prompts Iniciais

Fonte: Autoria própria.

Este estudo baseou-se em abordagens metodológicas científicas para sua construção e execução. A utilização de métodos consolidados na pesquisa científica é fundamental para trazer mais credibilidade e qualidade ao estudo realizado, pois os meios de desenvolvimento aplicados são amplamente reconhecidos e utilizados no meio acadêmico. Dessa forma, a execução deste trabalho envolveu a aplicação de diversas técnicas, que serão detalhadas e explicadas ao longo deste capítulo.

A adoção de uma metodologia científica rigorosa garante que os processos de coleta, análise e interpretação dos dados sejam conduzidos de maneira sistemática e objetiva. Isso não apenas fortalece a validade dos resultados obtidos, mas também facilita a reprodução do estudo por outros pesquisadores, contribuindo para a construção de um conhecimento mais robusto e confiável na área investigada.

No decorrer deste capítulo, serão apresentadas as etapas específicas do método científico utilizado, desde a revisão bibliográfica até o desenvolvimento experimental, coleta de dados, análise e interpretação dos resultados. Cada uma dessas etapas será descrita em detalhes, evidenciando como as práticas científicas foram aplicadas para assegurar a qualidade e integridade do estudo.

Esta pesquisa apresenta uma abordagem quantitativa que visa trazer um resultado de uma abordagem utilizada na comparação semântica textual entre frases.

4.1 Pesquisa Bibliográfica

Na realização de uma pesquisa científica, um dos passos primordiais é a realização da pesquisa bibliográfica. Essa abordagem dá ao pesquisador um panorama importante para que a escrita seja realizada de forma fundamentada. Através da pesquisa bibliográfica, é possível perceber, em um âmbito mais generalizado, o que há de existente sobre o assunto abordado na investigação. Neste caso, o estudo é voltado à utilização de Grandes Modelos de Linguagem na atividade de comparação semântica textual de frases em língua portuguesa brasileira, uma importante divisão oriunda da área de Processamento de Linguagem Natural.

Conforme afirma (FONSECA, 2002), "qualquer trabalho científico inicia-se com uma pesquisa bibliográfica, que permite ao pesquisador conhecer o que já se estudou sobre o assunto." Além disso, (FONSECA, 2002) descreve que "A pesquisa bibliográfica é feita a partir do levantamento de referências teóricas já analisadas, e publicadas por meios escritos e eletrônicos, como livros, artigos científicos e páginas de websites." Neste contexto, este trabalho utilizou como base, fundamentalmente, menções a artigos científicos publicados e de conhecimento consolidado.

A pesquisa bibliográfica não só orienta o pesquisador sobre os estudos já realizados na área, como também auxilia na identificação de lacunas no conhecimento, possibilitando a proposição de novas hipóteses e a construção de um referencial teórico robusto. Dessa forma, torna-se possível contextualizar o estudo em um panorama mais amplo, mostrando como ele se relaciona com os trabalhos existentes e contribuindo para o avanço da ciência.

Para a realização deste trabalho, foram consultados diversos artigos científicos, livros e fontes eletrônicas, que abordam tanto a teoria quanto as aplicações práticas dos Grandes Modelos de Linguagem e do Processamento de Linguagem Natural. As referências escolhidas foram cuidadosamente selecionadas para garantir que fossem de alta relevância e qualidade, permitindo um aprofundamento adequado no tema e proporcionando uma base sólida para as análises e discussões apresentadas.

Portanto, a pesquisa bibliográfica desempenha um papel crucial na estruturação de qualquer trabalho científico, pois oferece ao pesquisador uma compreensão profunda do estado da arte e das direções futuras que podem ser exploradas. A partir dessa base, este estudo busca contribuir significativamente para o campo do Processamento de Linguagem Natural, especialmente na comparação semântica textual de frases em língua portuguesa

brasileira.

4.2 Desenvolvimento Experimental

Para a realização desta pesquisa, foi utilizada uma abordagem de desenvolvimento experimental, focada no uso de grandes modelos de linguagem na atividade de comparação semântica textual entre frases em língua portuguesa brasileira. Os dados selecionados e o ambiente utilizado foram planejados minuciosamente para a execução eficiente de todas as etapas necessárias. A perspectiva experimental permite que o pesquisador realize tarefas complexas em ambientes controlados, garantindo a precisão e a confiabilidade dos resultados obtidos. A pesquisa experimental, segundo (GIL, 2007), "consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto".

Neste contexto, o desenvolvimento experimental foi essencial para uma observação direta e abrangente das relações de causa e efeito entre as variáveis envolvidas, como a estrutura dos prompts e a qualidade das respostas geradas pelos modelos de linguagem. A abordagem experimental também facilita a replicação do estudo, o que é fundamental para a verificação e validação dos resultados por outros pesquisadores. Além disso, a escolha cuidadosa dos dados propicia que os resultados sejam relevantes e aplicáveis ao contexto específico da pesquisa.

Em resumo, a metodologia experimental adotada nesta pesquisa não só fortaleceu a validade dos resultados obtidos, mas também contribuiu para a construção de um conhecimento científico robusto e confiável sobre o uso de grandes modelos de linguagem na comparação semântica textual em português brasileiro. A capacidade de controlar e manipular variáveis em um ambiente experimental proporcionou contribuições valiosas, oferecendo uma base sólida para futuras pesquisas e/ou aplicações práticas.

4.3 Base de Dados

A base de dados utilizada nesta pesquisa foi extraída do ASSIN II, a segunda Avaliação de Similaridade Semântica e Inferência Textual, um workshop realizado em conjunto com o STIL 2019. A Similaridade Semântica Textual quantifica o grau de

equivalência semântica entre duas sentenças. Os dados fornecidos pelo ASSIN II foram cuidadosamente anotados para garantir a qualidade e precisão das análises. O corpus de treino e validação compreende 6.500 e 500 pares de sentenças em português do Brasil, respectivamente, anotados tanto para inferência quanto para similaridade semântica. As anotações de similaridade semântica variam de 1 a 5. Além disso, o conjunto de teste inclui aproximadamente 3.000 pares de sentenças, garantindo a robustez dos modelos desenvolvidos e a consistência nas avaliações de desempenho. Nesta pesquisa foi utilizado o corpus composto por 500 pares de frases.

Utilizar a base de dados do ASSIN II proporciona uma fundação sólida para a comparação de similaridade semântica textual entre frases em língua portuguesa brasileira. A metodologia de avaliação seguiu rigorosamente as métricas estabelecidas, submetidos às métricas correlação de Pearson e Erro Quadrático Médio (MSE), conforme utilizados no evento, permitindo uma análise detalhada e precisa do desempenho do modelo. Os resultados obtidos nesta pesquisa não apenas refletem a eficácia do modelo proposto, mas também sublinham a importância de bases de dados bem estruturadas e anotadas, como as fornecidas pelo ASSIN II, para avanços significativos no campo do Processamento de Linguagem Natural.

5 Resultados e Discussões

Este capítulo tem por objetivo apresentar os resultados obtidos a partir da análise gerada por uma inteligência artificial generativa, especificamente o ChatGPT, na tarefa de comparação de similaridade semântica entre frases em língua portuguesa brasileira. Para uma análise abrangente dos resultados obtidos, as respostas geradas pela ferramenta foram avaliadas segundo duas métricas: Correlação de Pearson e Erro Quadrático Médio (MSE). Essas métricas foram escolhidas por terem sido utilizadas no workshop ASSIN II - "II Avaliação de Similaridade Semântica e Inferência Textual". O workshop, realizado em 2019, foi uma parceria com o STIL - Symposium in Information and Human Language Technology, e teve como objetivo desenvolver abordagens de inteligência artificial voltadas para a avaliação da similaridade semântica textual e o reconhecimento de inferência entre pares de frases em língua portuguesa brasileira.

Dessa forma, a presente pesquisa tomou como base os parâmetros estabelecidos no evento, focando especificamente na abordagem de similaridade semântica textual por meio das métricas citadas anteriormente, as quais serviram de mediadoras para os resultados obtidos no workshop.

5.1 Construção e Análise de Prompts

A construção dos prompts utilizados na pesquisa seguiu etapas fundamentais, demonstrando a importância do refinamento e alinhamento da redação de cada prompt. Abaixo, descrevemos as etapas desenvolvidas para a criação dos prompts utilizados neste trabalho.

As etapas de utilização e teste dos prompts ocorreram de duas formas. Na primeira, os prompts foram submetidos diretamente à ferramenta do ChatGPT disponível no site da OpenAI. Na segunda abordagem, os prompts foram submetidos ao modelo via código Python, utilizando a biblioteca LangChain. Para essa utilização, é necessário o uso de uma API Key da OpenAI, que permite o uso do modelo de forma "indireta", através de código-fonte. Essa divisão de etapas se deu devido a uma das limitações desta pesquisa: o uso da API Key da OpenAI possui custos de utilização, e nas etapas iniciais da pesquisa ainda não havia verba disponível para testes diretos via LangChain.

Para a construção dos prompts, observamos as orientações provenientes da Engenharia de Prompt. Utilizamos duas abordagens: Prompts Zero-Shot e Prompts Few-Shot. Prompts Zero-Shot são redigidos sem nenhum exemplo ao modelo, enquanto em Prompts Few-Shot alguns modelos de resposta são fornecidos ao modelo, permitindo que a resposta gerada tenha como base uma lista de respostas anteriores, o que pode propiciar maior assertividade por parte da IA.

Todos os prompts foram criados em cinco níveis, conforme reflexões levantadas por (SANTU; FENG, 2023), nomeados aqui como Nível 01, Nível 02, Nível 03, Nível 04 e Nível 05. Os prompts de Nível 01 foram redigidos com comandos diretos ao modelo, variando apenas que no exemplo de Nível 01 do Few-Shot, alguns exemplos foram inseridos junto ao prompt. Os prompts do Nível 02 até o Nível 05 já utilizaram mais especificações em sua redação, que foram evoluindo e aumentando gradativamente com o avanço de níveis do prompt. Abaixo, podemos perceber a evolução básica entre um Prompt Nível 01 e um Prompt Nível 05:

Redação do Prompt Nível 01 - Técnica Zero-Shot.

> Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.
> 1.0, 1.25, 1.50 ou 1.75: De ”As frases são completamente diferentes.”até ”As frases possuem pouca similaridade.”
> 2.0, 2.25, 2.50 ou 2.75: De ”As frases possuem pouca similaridade.”até ”As frases possuem similaridades.”
> 3.0, 3.25, 3.50 ou 3.75: De ”As frases possuem similaridades.”até ”As frases possuem muita similaridade.”
> 4.0, 4.25, 4.50 ou 4.75: De ”As frases possuem muita similaridade.”até ”As frases iguais ou são extremamente similares.”
> 5.0: As frases iguais ou são extremamente similares.
> Frase A: frase1
> Frase B: frase2
>
> Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.

Redação do Prompt Nível 05 - Técnica Zero-Shot.

Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.
1.0, 1.25, 1.50 ou 1.75: De "As frases são completamente diferentes."até "As frases possuem pouca similaridade."
2.0, 2.25, 2.50 ou 2.75: De "As frases possuem pouca similaridade."até "As frases possuem similaridades."
3.0, 3.25, 3.50 ou 3.75: De "As frases possuem similaridades."até "As frases possuem muita similaridade."
4.0, 4.25, 4.50 ou 4.75: De "As frases possuem muita similaridade."até "As frases iguais ou são extremamente similares."
5.0: As frases iguais ou são extremamente similares.

***
Geralmente, frases com maior índice de similaridade possuem:
1º: Sintaxe e/ou a estrutura das palavras similares ou iguais.
2º: As frases possuem semântica e/ou o significado de cada palavra contidas nelas são similares ou iguais.
3º: Crie palavras-chave e termos relevantes para cada frase. As frases possuem palavras-chave e/ou termos relevantes similares ou iguais.
4º: As frases possuem contexto e/ou intenção similares ou iguais;
5º: As frases possuem palavras que são sinônimos.
***

Frase A: frase1
Frase B: frase2

Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.

Como tarefa inicial para escolher o prompt ideal para a realização da análise e comparação dos dados, foram gerados cinco níveis de prompt com a técnica Zero-Shot e outros cinco com a técnica Few-Shot. De posse dos dez prompts iniciais, foram escolhidos aleatoriamente 10 pares de frases do dataset utilizado. Descritas abaixo:

FRASES 01 - Índice no Dataset: 2501
Valor de Similaridade Anotado: 3.50
Frase 01: De acordo com o relatório, foram notificados 6.052 casos suspeitos de dengue, sendo 641 descartados.

Frase 02: Do total de casos notificados, 10.768 foram confirmados como dengue e 15.202 descartados.

FRASES 02 - Índice no Dataset: 2514
Valor de Similaridade Anotado: 1.25
Frase 01: A previsão para a taxa de câmbio em 2015 ficou em R$ 3,20.
Frase 02: Para 2016, a previsão de superávit comercial permaneceu em US$ 9,95 bilhões.

FRASES 03 - Índice no Dataset: 2519
Valor de Similaridade Anotado: 2.50
Frase 01: As exportações somaram US$ 57,931 bilhões nos primeiros quatro meses deste ano e as importações totalizaram US$ 62,997 bilhões.
Frase 02: As importações totalizaram US$ 19,218 bilhões.

FRASES 04 - Índice no Dataset: 2503
Valor de Similaridade Anotado: 3.00
Frase 01: Em Campinas, no interior do estado, já são mais de 30 mil casos confirmados de dengue.
Frase 02: São 333 municípios com notificações da doença e 246 têm confirmados casos de dengue.

FRASES 05 - Índice no Dataset: 2659
Valor de Similaridade Anotado: 4.75
Frase 01: Os demais agentes públicos serão alocados na classe econômica.
Frase 02: Todo o resto dos funcionários públicos terá que embarcar na classe econômica.

FRASES 06 - Índice no Dataset: 2652
Valor de Similaridade Anotado: 1.75
Frase 01: O consagrado profissional da dramaturgia aparece ao lado de sua esposa, Marilene Saade.
Frase 02: Em uma das imagens, Marilene Saade aparece com um celular na mão.

FRASES 07 - Índice no Dataset: 2682
Valor de Similaridade Anotado: 5.00
Frase 01: A quadra teve 2.143 apostas ganhadoras, que levaram prêmio de R$ 6.133,01 cada uma.
Frase 02: Outras 2.143 apostas acertaram quatro números e levaram R$ 6.133,01

FRASES 08 - Índice no Dataset: 2539
Valor de Similaridade Anotado: 4.00
Frase 01: Em maio deste ano, eles reataram, mas terminaram novamente no dia 24 de setembro.
Frase 02: Em maio deste ano, reataram, voltando a terminar em setembro.

FRASES 09 - Índice no Dataset: 2557
Valor de Similaridade Anotado: 1.00
Frase 01: Eu falo aqui no programa, e estou à disposição.
Frase 02: Gilmar Rinaldi desafiou o senador Romário.

FRASES 10 - Índice no Dataset: 2526
Valor de Similaridade Anotado: 3.25
Frase 01: A Caixa Econômica Federal faz os sorteios da Mega-Sena duas vezes por semana, às quartas-feiras e aos sábados
Frase 02: A Mega-Sena terá três concursos nesta semana, segundo a Caixa Econômica Federal.

Dessa forma, todos os dez tipos de prompts foram submetidos ao ChatGPT através do site da ferramenta. De posse dos resultados gerados, conforme a tabela abaixo, observamos qual nível e qual técnica foram mais efetivos na precisão dos valores. Outro indicador observado foi a orientação descrita por (KOJIMA et al., 2022). Eles afirmam que utilizar a frase "Execute etapa por etapa" ao final de cada prompt melhora substancialmente a resposta gerada pelo modelo. Assim, analisamos três indicadores na escolha dos prompts finais: 1º Uso da técnica da Engenharia de Prompt, 2º Nível de detalhamento da redação do prompt, e 3º Utilização da abordagem de (SANTU; FENG, 2023) ao analisar prompts com ou sem o uso da frase em seu final. Vale salientar que todos os prompts submetidos ao modelo passaram por refinamentos em busca de um melhor resultado; as mudanças feitas em cada prompt constam nos apêndices desta pesquisa.

Com os resultados da análise inicial dos Prompts, foi percebido nesta abordagem experimental que os Prompts de Nível 01, sem a utilização das frases "Execute etapa por etapa" e sem um maior refinamento em sua redação, obtiveram melhores resultados em comparação com as demais abordagens. De posse dos prompts de melhor resultado, foi possível realizar a análise utilizando LangChain. Assim, todos os cinco níveis de prompt

da abordagem Zero-Shot foram analisados.

Para a escolha dos prompts utilizados, bem como de qual das abordagens oriundas da Engenharia de Prompt utilizar, foi realizado uma avaliação inicial com os modelos e os seus respectivos resultados na avaliação de similaridade semântica foram cruciais para a tomada de decisão

Na investigação final, os modelos de prompts foram avaliados pela inteligência artificial através da biblioteca Python LangChain. Cada nível de prompt foi avaliado com cada um dos 500 pares de frases que constam no dataset. Os resultados obtidos foram discutidos em duas abordagens. Primeiro, foi feita uma análise dos valores gerados pelas métricas Correlação de Pearson e Erro Médio Quadrático (MSE), métricas utilizadas no Workshop ASSIN II. Em seguida, foi realizada uma comparação entre os resultados desta pesquisa, através das métricas citadas, com os resultados do evento ASSIN II. Obteve-se como resultado final que o Prompt 01 destacou-se e atingiu melhores resultados nas métricas analisadas, além de ter obtido um desempenho bom na comparação com os resultados dos participantes do ASSIN II.

5.2 Desempenho dos Prompts

Tabela 2 – Resultados Prompts Nível 01

Frase	Zero-Shot A	Few-Shot A	Zero-Shot B *	Few-Shot B *
2514	1.0	1.0	1.5	1.0
2519	3.0	2.0	3.0	2.25
2503	3.5	2.75	3.5	2.25
2659	4.5	3.25	4.0	3.25
2652	2.5	1.25	2.0	2.25
2682	5.0	3.25	4.5	4.25
2531	4.0	1.75	4.0	2.25
2526	3.5	2.25	2.5	2.25
2501	3.5	3.0	3.0 9	3.0
2557	1.25	1.75	2.0	1.5

Fonte: Autoria própria.

* As colunas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Tabela 3 – Resultados Prompts Nível 02

Frase	Zero-Shot A	Few-Shot A	Zero-Shot B *	Few-Shot B *
2514	1.0	1.75	1.0	1.0
2519	2.5	1.75	3.5	2.75
2503	1.5	2.75	2.0	1.0
2659	3.75	2.75	4.0	4.0
2652	1.0	2.25	1.0	2.75
2682	3.75	2.25	4.25	2.75
2531	4.25	2.25	4.0	2.75
2526	1.5	2.25	1.0	2.75
2501	3.0	2.5	2.25	2.5
2557	2.0	1.5	1.5	1.75

Fonte: Autoria própria.

* As colunas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Tabela 4 – Resultados Prompts Nível 03

Frase	Zero-Shot A	Few-Shot A	Zero-Shot B *	Few-Shot B *
2514	1.5	1.0	1.0	1.0
2519	1.0	2.5	3.0	1.0
2503	1.5	1.0	2.0	1.0
2659	4.25	3.5	4.5	1.75
2652	1.5	1.0	1.0	1.0
2682	4.25	1.0	4.5	1.0
2531	4.25	3.5	3.5	3.75
2526	1.5	1.0	1.5	1.0
2501	3.75	3.0	3.0	3.0
2557	1.5	1.75	1.75	1.75

Fonte: Autoria própria.

* As colunas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Tabela 5 – Resultados Prompts Nível 04

Frase	Zero-Shot A	Few-Shot A	Zero-Shot B *	Few-Shot B *
2514	1.5	1.0	1.5	1.25
2519	2.0	1.0	2.5	1.25
2503	2.0	1.0	2.5	1.25
2659	3.5	4.0	3.75	2.25
2652	1.5	1.0	1.25	1.25
2682	4.25	4.0	3.5	1.25
2531	4.25	3.5	3.5	1.25
2526	2.5	1.0	2.0	1.25
2501	3.5	3.25	3.5	3.0
2557	1.25	1.5	1.25	1.75

Fonte: Autoria própria.

* As colunas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Tabela 6 – Resultados Prompts Nível 05

Frase	Zero-Shot A	Few-Shot A	Zero-Shot B *	Few-Shot B *
2514	1.0	1.0	1.5	2.25
2519	1.0	1.0	1.0	1.75
2503	1.0	1.25	1.0	1.25
2659	4.0	3.25	4.0	1.25
2652	1.0	1.25	1.0	1.0
2682	3.0	1.5	4.5	1.5
2531	4.0	2.5	3.5	2.0
2526	1.0	1.25	2.0	1.25
2501	3.0	2.75	3.25	3.0
2557	1.5	1.75	1.75	1.75

Fonte: Autoria própria.

* As colunas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Diante dos resultados expressos nas tabela 2, 3, 4, 5 e 6, foi analisado qual técnica melhor obteve resultados se comparados com os valores anotados no dataset. Para a realização desta análise foi considerado duas perspectivas: O prompt que mais obteve resultados idênticos aos valores anotados no dataset e o prompt que mais se aproximou desses valores. Sendo assim, obteve-se os seguintes resultados:

Tabela 7 – Resultados com Maior Proximidade dos Valores Anotados

Técnica	Nível 01	Nível 02	Nível 03	Nível 04	Nível 05	Total
Zero-Shot	7	3	6	6	4	26
Few-Shot	2	3	2	1	3	11
Zero-Shot*	3	5	5	3	5	21
Few-Shot*	2	3	2	1	2	10

Fonte: Autoria própria.
* As linhas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Tabela 8 – Resultados Iguais aos Valores Anotados

Técnica	Nível 01	Nível 02	Nível 03	Nível 04	Nível 05	Total
Zero-Shot	3	1	0	1	1	6
Few-Shot	0	0	1	0	0	1
Zero-Shot*	1	1	0	1	0	3
Few-Shot*	0	0	0	1	0	1

Fonte: Autoria própria.
* As linhas 4 e 5 utilizaram a abordagem “Execute etapa por etapa” em sua redação.

Dessa forma, observando os resultados de cada técnica e cada nível de prompt, verificou-se que o Prompt Nível 01 obteve melhor desempenho em comparação com seus pares, alcançando uma assertividade na proximidade de valores por 26 vezes e gerando resultados idênticos aos anotados em 6 ocasiões.

Com os dados obtidos pelos métodos adotados, decidiu-se utilizar a abordagem Zero-Shot, sem o uso da expressão ”Execute etapa por etapa”. Para um estudo mais aprofundado do caso e visando favorecer a possibilidade de discussão do experimento, optou-se por realizar a abordagem experimental com todos os cinco níveis de prompts.

Essa escolha metodológica se deu em virtude da análise detalhada dos resultados

iniciais, onde se constatou que o Prompt Nível 01, sem o comando adicional, apresentava uma maior consistência e precisão. Além disso, a aplicação dos cinco níveis de prompts permitiu uma avaliação mais robusta e abrangente, proporcionando uma compreensão mais completa do comportamento do modelo em diferentes contextos de complexidade.

A abordagem experimental foi, portanto, fundamentada na necessidade de explorar de forma abrangente os diferentes níveis de especificação e detalhamento dos prompts, buscando identificar quais variações poderiam proporcionar melhorias nos resultados de similaridade semântica textual. Essa estratégia permitiu não apenas a validação dos resultados preliminares, mas também a identificação de padrões e tendências que podem contribuir para futuras pesquisas na área de Processamento de Linguagem Natural (PLN). A análise minuciosa dos dados coletados evidenciou a importância de uma abordagem meticulosa e bem estruturada na elaboração de prompts, destacando a relevância da engenharia de prompt na obtenção de resultados precisos e consistentes.

Ao final, os resultados mostraram que o Prompt Nível 01 se destacou não apenas em termos de precisão, mas também em sua capacidade de gerar respostas consistentes e alinhadas com as anotações manuais, corroborando a eficácia da abordagem adotada e proporcionando contribuições valiosas para o aprimoramento das técnicas de engenharia de prompt em aplicações de PLN.

5.3 Resultados da Correlação de Pearson

A métrica de Correlação de Pearson visa indicar como duas variáveis se relacionam entre si, determinando se estão fortemente associadas, pouco associadas ou se não há associação entre elas. Esta métrica é essencial para compreender a força e a direção da relação linear entre as variáveis analisadas, sendo amplamente utilizada em estudos de correlação por sua robustez e simplicidade interpretativa.

A Correlação de Pearson r é calculada a partir da seguinte formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{5.1}$$

Onde:

- x_i e y_i são os valores das variáveis.

- $\bar{x}$ e $\bar{y}$ são as médias das variáveis analisadas x e y.
- $\sum$ é a soma sobre todos os valores.

A formula da Correlação de Pearson gera um resultado de correlação entre as variáveis analisadas, esse resultado tem uma variação que vai desde -1 a 1, que podemos interpretar da seguinte forma:

- Resultado 1: Quanto mais próximo do valor 1, há indicação de que a correlação é maior, onde o resultado 1 indica uma correlação linear positiva perfeita.
- Resultado -1: Quanto mais próximo de -1, o resultado indica uma menor relação entre as variáveis. Onde, especificamente, -1 indica uma correlação linear negativa perfeita.
- Resultado 0: Já se o valor do resultado for 0, demonstra que não há correlação linear.

No caso da métrica em questão, podemos indicar que no contexto da pesquisa, quanto mais próximo de 1 os resultados estiverem, melhor é o desempenho da abordagem utilizada, onde o contrário indica justamente o oposto, que o modelo utilizado não obteve um bom resultado na tarefa de comparação de similaridade semântica entre as frases se compararmos com os resultados anotados do dataset "Y".

A outra métrica utilizada foi o Erro Quadrático Médio (MSE). Essa métrica avalia a média dos quadrados dos erros preditos pelo modelo de inteligência artificial utilizado. Observa-se a diferença entre os valores preditos pelo modelo "$\hat{y}$" e os valores reais anotados na base de dados "Y", através da seguinte fórmula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.2}$$

A fórmula do Erro Quadrático Médio (MSE) gera valores que são analisados da seguinte forma: Quanto menor o valor obtido, melhor é a precisão do modelo utilizado. Porém, quanto maior forem os valores resultantes, menor é a assertividade da predição em contexto.

A base de dados utilizada no experimento foi disponibilizada no site do Workshop ASSIN II, onde continua acessível para livre consulta e uso. Esta base é composta por três conjuntos distintos. Os dois primeiros conjuntos são destinados ao treinamento e validação do modelo, contendo respectivamente 6.500 e 500 pares de frases em língua

portuguesa do Brasil. Além disso, há um terceiro conjunto utilizado para teste, composto por 3.000 pares de sentenças. Todos os dados mencionados foram anotados manualmente. Para o desenvolvimento desta pesquisa, especificamente, utilizou-se o conjunto composto por 500 pares de frases.

Essas bases de dados são fundamentais para avaliar e treinar modelos de inteligência artificial voltados à tarefa de avaliação de similaridade semântica textual e inferência textual. A anotação manual assegura a qualidade e a consistência dos dados, essenciais para a confiabilidade dos resultados obtidos na pesquisa.

O acesso aberto a esses dados promove a replicabilidade e a transparência dos experimentos científicos, permitindo que outros pesquisadores possam validar e comparar métodos e resultados. O uso específico do conjunto de 500 pares de frases neste estudo ressalta a importância da escolha criteriosa dos dados de acordo com os objetivos específicos de cada pesquisa.

Para a avaliação do modelo utilizado, foram empregadas cinco abordagens diferentes de prompts. Cada uma delas tinha como objetivo orientar o ChatGPT na avaliação da similaridade textual entre os pares de frases. As diretrizes provenientes da engenharia de prompts foram utilizadas para selecionar os prompts mais adequados. Dessa forma, os prompts foram categorizados por nível (1 a 5), sendo que quanto maior o nível do prompt, maior e mais complexo é o conjunto de instruções fornecido à IA generativa em questão.

Para a análise dos resultados obtidos, foram gerados gráficos e tabelas que facilitarão o entendimento acerca do desempenho do modelo utilizado.

Tabela 9 – Resultado - Correlação de Pearson.

Prompt Level	Result
Level 01	0.640
Level 02	0.404
Level 03	0.426
Level 04	0.444
Level 05	0.398

Fonte: Autoria própria.

A Tabela 1 apresenta os resultados relacionados a cada prompt utilizado durante a pesquisa, especificamente na métrica Correlação de Pearson. Observa-se que todos os resultados foram positivos, ou seja, acima de zero. Esse fator demonstra que todos os prompts utilizados tiveram correlação direta com as respostas anotadas no dataset, embora tenha havido uma perda de correlação à medida que aumentava o nível de detalhamento oferecido ao modelo de inteligência artificial.

Conforme mostrado na tabela, o prompt Level 01 obteve melhor desempenho em comparação aos seus pares. Esse prompt tinha instruções diretas a serem seguidas, sem muitos detalhamentos adicionais para o grande modelo de linguagem. Os demais prompts apresentaram resultados inferiores ao primeiro, com uma leve oscilação à medida que o nível de detalhamento aumentava. Um aspecto interessante foi observado nos resultados dos prompts Level 04 e 05. Esperava-se que o resultado crescesse de um para o outro, dado o maior detalhamento do Level 05, mas na verdade houve uma diminuição no desempenho. A fim de ilustrar melhor os resultados, observamos a Matriz de Correlação ente os Prompts:

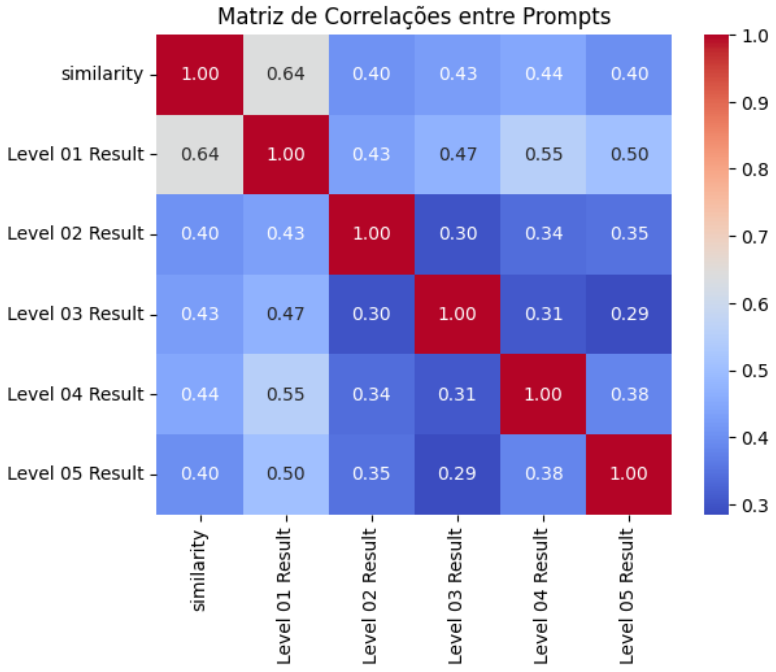

Figura 1 – Matriz de Correlação
Fonte: Autoria própria.

A Figura 1 demonstra com clareza os resultados obtidos a partir da métrica de Correlação de Pearson. Observa-se, mais uma vez, que o Prompt Level 1 destaca-se dos demais por apresentar um grau maior de correlação com os valores anotados. É

perceptível também que os resultados dos demais prompts não diferem significativamente entre si, sendo o segundo melhor resultado 0,44 e o pior resultado 0,40.

5.4 Erro Quadrático Médio (MSE)

Tabela 10 – Resultado - Erro Quadrático Médio (MSE).

Prompt Level	Result
Level 01	0.468
Level 02	0.697
Level 03	0.682
Level 04	0.654
Level 05	0.697

Fonte: Autoria própria.

Observa-se na Tabela 2 os valores resultantes da avaliação utilizando o Erro Quadrático Médio (MSE). Assim como observado na métrica anterior, o Prompt Level 1 obteve o melhor valor dentre os resultados em comparação com os demais prompts analisados. Valores menores resultantes do MSE indicam melhores predições do modelo em comparação com os resultados reais de similaridade anotados no dataset. Dessa forma, o Prompt Level 1, mesmo sendo o que não utilizou princípios avançados da Engenharia de Prompt, foi o que obteve as melhores predições. Esse prompt utilizou o seguinte comando direto para o modelo: “Sua tarefa é indicar um valor de similaridade semântica entre a Frase A e a Frase B”, seguido pela apresentação de cada par de frases a ser comparada.

Os Prompts 2 e 5 obtiveram valores iguais em seus resultados, 0,697. Esse valor pode ser considerado ruim em comparação com o resultado do Prompt 1, além de terem sido os piores resultados obtidos a partir da métrica utilizada. Ao analisar o desempenho desses prompts, percebe-se a necessidade de ajustes em seu desenvolvimento. Possivelmente, eles introduziram complexidade desnecessária ou até informações insuficientes ao modelo para obter predições mais precisas.

Os Prompts 3 e 4 obtiveram resultados de 0,682 e 0,654, respectivamente. Esses valores também são considerados altos, tendo em vista que o menor valor foi 0,468. Isso

indica a necessidade de refinamento dos prompts em busca de melhores resultados.

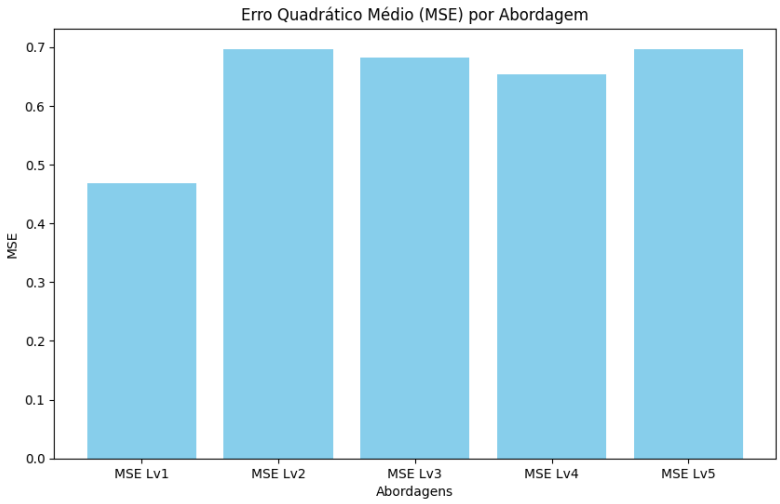

Figura 2 – Resultados Erro Quadrático Médio
Fonte: Autoria própria.

A Figura 2 demonstra graficamente os resultados obtidos. Similarmente ao resultado da métrica anterior, percebemos aqui uma maior diferença entre os valores analisados. O Prompt Level 1, com um resultado de 0,465, destaca-se positivamente.Os demais prompts continuaram a apresentar pouca diferença entre seus resultados, com uma variação mínima de 0,043 entre o pior e o segundo melhor resultado.

Os Prompts 2, 3, 4 e 5, apesar de utilizarem maior detalhamento de informações e comandos a serem seguidos, não conseguiram superar a simplicidade e eficácia do Prompt 1.

Essa análise evidencia a importância de um equilíbrio na engenharia de prompts. O excesso de detalhamento ou complexidade pode, de certa forma, reduzir a precisão das predições do modelo. Portanto, embora a engenharia de prompt seja uma ferramenta poderosa para orientar os modelos de linguagem, é essencial que o seu uso seja feito de forma cuidadosa, evitando a introdução de elementos que possam confundir o modelo em vez de guiá-lo de forma clara.

5.5 Resultados da Pesquisa X Resultados do ASSIN II

Será apresentada uma análise comparativa dos resultados obtidos nesta pesquisa em relação aos resultados dos grupos de pesquisadores que participaram do ASSIN II. O evento contou com a participação de dez grupos, cada um desenvolvendo três

propostas para a avaliação de similaridade semântica e inferência textual. Esta pesquisa avaliou, especificamente, o cenário de similaridade semântica, cujos resultados serão apresentados a seguir. Os gráficos abaixo sintetizam os valores obtidos pelos diferentes grupos participantes do workshop, além dos resultados das abordagens desenvolvidas durante esta pesquisa. Isso nos permite realizar uma comparação direta da efetividade das propostas.

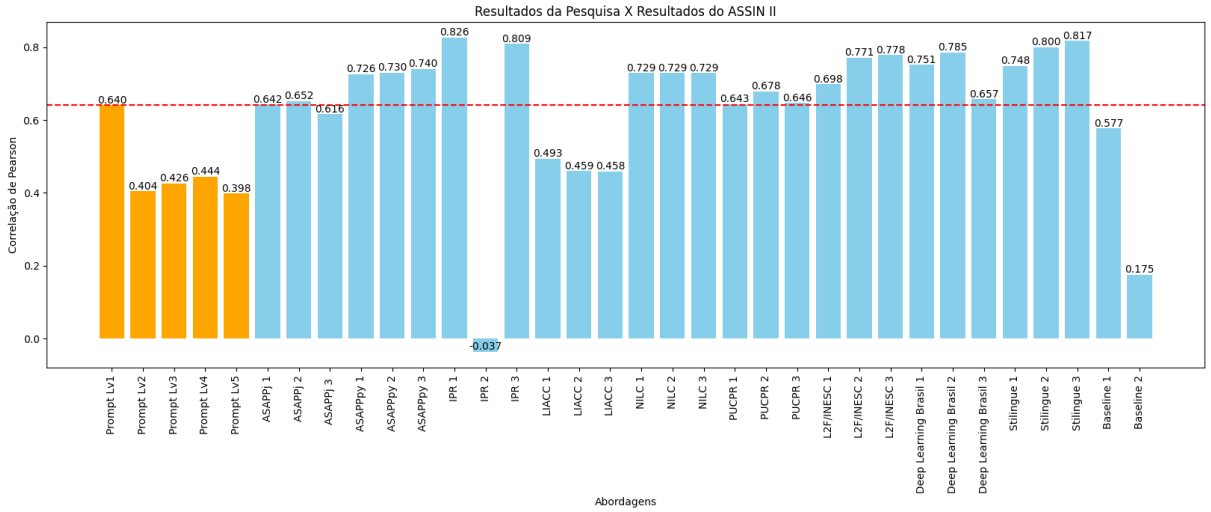

Figura 3 – Resultados em Correlação de Pearson
Fonte: Autoria própria.

A Figura 3 apresenta os resultados avaliados a partir da métrica de Correlação de Pearson. Nela é possível perceber os comparativos entre as cinco abordagens experimentais e os valores resultantes dos competidores do ASSIN II. Utilizando como base a abordagem com melhor resultado obtido, referenciamos os valores do Prompt 1. Ao analisar os números, nota-se que o Prompt 1 obteve apenas a 23ª melhor colocação em relação aos resultados do ASSIN II. Esta colocação demonstra claramente a necessidade de ajustes nas técnicas empregadas durante este trabalho e reforça a importância de melhorias e refinamentos nos prompt em busca de atingir um melhor resultado.

A Figura 4 ilustra o gráfico comparativo dos resultados do Erro Quadrático Médio (MSE). Nesta comparação, o resultado do Prompt 1 obteve uma excelente posição, alcançando a 2ª melhor colocação em comparação com todas as demais abordagens. Apenas o experimento 2 do grupo “Stilingue” obteve uma melhor colocação, com um MSE igual a 0,390, enquanto o Prompt 1 obteve um valor de 0,468. Torna-se assim um excelente resultado para a abordagem proposta por este trabalho.

Assim, na métrica de Correlação de Pearson, o comparativo com o resultado do

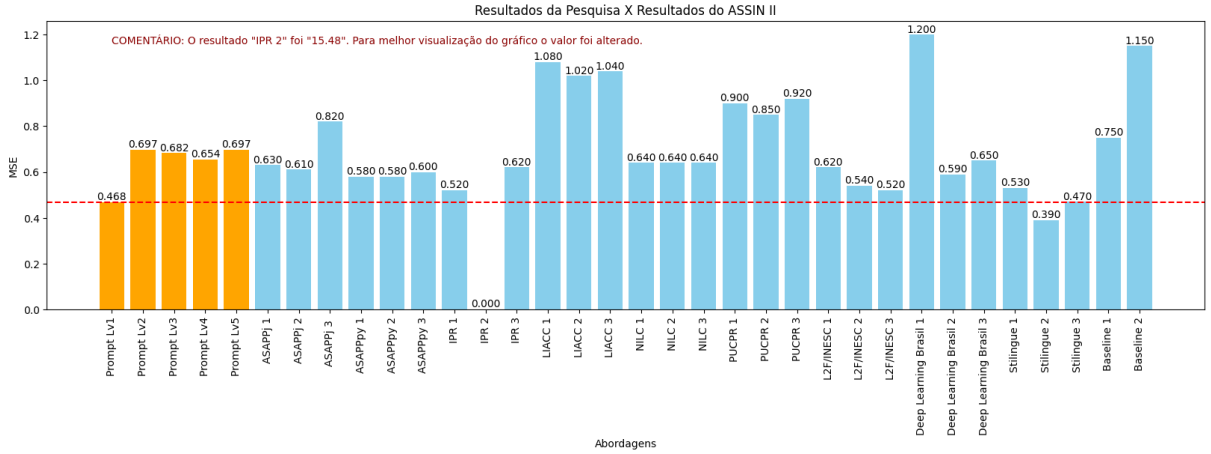

Figura 4 – Resultados em Erro Quadrático Médio - MSE
Fonte: Autoria própria.

Prompt 1 evidencia que 23% dos resultados do ASSIN II necessitam de melhorias na performance para identificar corretamente a similaridade semântica. Considerando que a Correlação de Pearson foi a métrica primária utilizada no ASSIN II, evidencia-se a necessidade real de melhorias nos resultados gerais.

Ao discutir os resultados do Erro Quadrático Médio, o desempenho do Prompt 1 demonstra que 97% dos resultados dos grupos durante o evento necessitam de melhorias e ajustes para reduzir a taxa de erros nas predições utilizadas nesta abordagem. A superioridade do Prompt 1 em MSE mostra que ele é altamente eficiente em realizar predições precisas de similaridade semântica, minimizando os erros de forma significativa em relação à maioria dos modelos concorrentes. Essa competência em MSE é indicativa de que as técnicas e abordagens utilizadas no desenvolvimento do experimento são sólidas e eficazes, justificando sua aplicação e promovendo futuras investigações nesta linha de pesquisa, mesmo que necessitem de melhorias visando atingir um melhor resultado na métrica Correlação de Pearson.

Dessa forma, a análise dos resultados através das métricas de Correlação de Pearson e MSE evidencia áreas cruciais para aprimoramento em seus prompts. Análises como essa são fundamentais para o avanço contínuo da precisão e eficiência dos modelos de similaridade semântica, promovendo uma base sólida para futuras pesquisas e desenvolvimento na área.

5.6 Discussões

Tomando como base os resultados obtidos a partir das análises realizados, pode-se observar alguns pontos importantes, tais quais:

1. Tratando-se dos resultados relacionados à variedade de prompts e ao uso de engenharia de prompt, percebe-se um distanciamento entre o que é destacado pela literatura e os resultados obtidos nesta abordagem específica de comparação de similaridade textual. O padrão dos resultados sugere que, embora a complexidade adicional fornecida pelas orientações advindas da Engenharia de Prompt possa oferecer mais informações ao modelo, ela também pode introduzir variações na qualidade das predições geradas. O comportamento da LLM pode ser impactado até mesmo pela menor mudança na redação do prompt utilizado, tornando-se sensível a mudanças e propenso a fornecer respostas mais variadas, em comparação ao que se almeja. Contrariando a expectativa de que o aumento de funções e a complexidade que compõem um prompt mais elaborado garantam maior qualidade no resultado gerado, observa-se que isso nem sempre ocorre. Dessa forma, é de suma importância realizar um processo mais extenso de aprimoramento dos prompts utilizados, pois é possível notar que as respostas tendem a ser melhores e mais assertivas quando se utiliza a forma mais adequada de introduzir os comandos para o modelo.

   Os resultados dessa etapa da pesquisa são valiosos e importantes para observamos duas perspectivas: primeiramente, que a Engenharia de Prompt necessita de constante aprimoramento científico, pois ainda é uma área iniciante no contexto acadêmico. Portanto, é necessário fomentar estudos científicos sobre a área, gerando um maior aprofundamento dos conhecimentos e permitindo melhores orientações para seus usuários. Em seguida, é importante mencionar que o refinamento dos prompts utilizados na execução de comandos para uma inteligência artificial é vital para a obtenção de melhores respostas, mesmo que seja um trabalho possivelmente complexo até que os ajustes resultem em um prompt ideal para a finalidade de avaliação de similaridade semântica textual, algo tão importante na área de Processamento de Linguagem Natural. Esse pensamento corrobora com o que é dito por (NASCIMENTO, 2024) em seu estudo: ”A formulação do prompt influencia quais padrões e conhecimentos são ativados no modelo [...]. Mesmo pequenas variações na formulação podem levar a interpretações e construções de contexto

diferentes."

Ampliando essa análise, observa-se que a introdução de complexidade adicional através de prompts mais detalhados pode, paradoxalmente, levar a uma degradação na qualidade das predições. Isso se deve à sensibilidade do modelo a pequenas variações na formulação dos prompts, o que pode desencadear uma série de respostas menos consistentes. Portanto, a expectativa inicial de que a complexidade e a especificidade adicionais melhorariam a qualidade das respostas não foi confirmada nesta pesquisa. Ao contrário, prompts mais simples e diretos mostraram-se frequentemente mais eficazes.

Consequentemente, a engenharia de prompt se revela como um campo que exige refinamento contínuo e meticuloso. Isso implica que os pesquisadores e profissionais da área devem investir tempo e esforço significativos para entender melhor como formular prompts de maneira que otimizem a resposta do modelo. Este processo de refinamento é fundamental para avançar na obtenção de resultados mais precisos e confiáveis, contribuindo para o desenvolvimento de técnicas mais eficazes no uso de modelos de linguagem natural.

Por fim, a importância de ajustar e aperfeiçoar os prompts não pode ser subestimada. Este trabalho reforça a necessidade de um contínuo desenvolvimento científico e tecnológico na área, garantindo que as ferramentas de Processamento de Linguagem Natural sejam cada vez mais precisas e úteis.

2. Ao analisar a comparação entre os resultados desta pesquisa e os resultados dos grupos participantes do Workshop ASSIN II, pode-se observar que houve uma boa classificação da proposta aqui discutida na métrica Erro Quadrático Médio (MSE) e uma classificação inferior ao compararmos com a métrica Correlação de Pearson. A abordagem aqui sugerida obteve o segundo melhor resultado na análise com Erro Quadrático Médio (MSE) e apenas a vigésima terceira melhor classificação na análise de Correlação de Pearson.

   De posse desses resultados, é válido afirmar, acerca das discrepâncias dos resultados em cada métrica, que os resultados desta pesquisa são promissores e reafirmam que os Grandes Modelos de Linguagem conseguem obter resultados tão eficazes quanto os dos humanos, independentemente das tarefas destinadas. Além disso, é importante considerar que a avaliação da similaridade semântica na língua

portuguesa brasileira apresenta desafios únicos devido às suas diversas nuances linguísticas, que podem levar a falsos positivos ou negativos na classificação dos resultados. Mesmo assim, a proposta mostrou-se promissora, destacando-se entre os melhores resultados em termos de MSE. No entanto, para melhorar a classificação na métrica de Correlação de Pearson, será necessário ajustar os modelos e talvez até reavaliar os prompts utilizados.

Sendo assim, ressalta-se que a formulação dos prompts e a natureza das tarefas podem impactar diferentemente as métricas. Prompts que fornecem informações mais detalhadas podem melhorar o MSE ao fornecer mais contexto ao modelo, mas podem não necessariamente alinhar as predições com a tendência linear dos valores reais, resultando em uma Correlação de Pearson mais baixa. A natureza e a variação das tarefas podem ter contribuído para essa discrepância entre as métricas utilizadas no ASSIN II e nesta pesquisa.

Portanto, é essencial continuar investindo em pesquisas que busquem aperfeiçoar esses modelos, tendo em vista que a precisão e a eficácia dos Grandes Modelos de Linguagem são fundamentais para diversas aplicações práticas. A continuidade dos estudos e a implementação de ajustes finos são passos necessários para alcançar um desempenho ainda mais robusto em ambas as métricas analisadas.

5.6.1 Resultados Anotados X Resultados Preditos

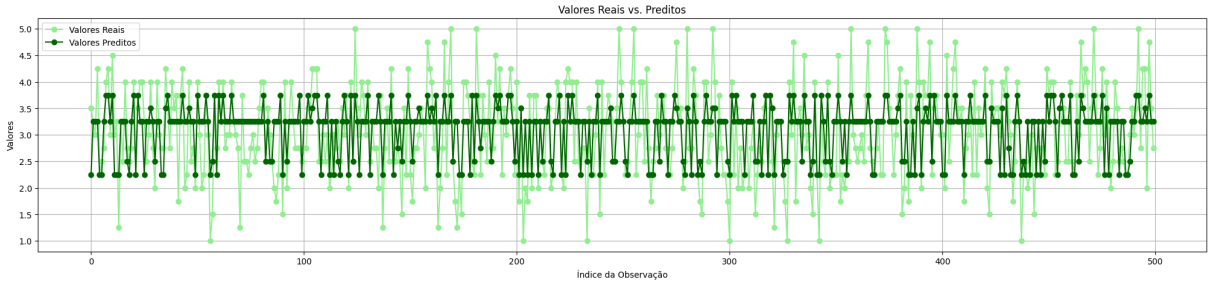

Figura 5 – Resultados Anotados X Resultados Preditos - Prompt Nível 01
Fonte: Autoria própria.

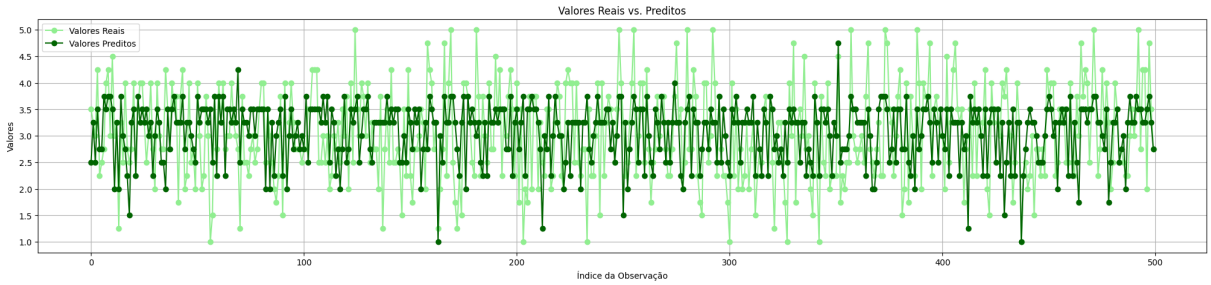

Figura 6 – Resultados Anotados X Resultados Preditos - Prompt Nível 05
Fonte: Autoria própria.

As figuras 5 e 6 ilustram a relação entre os valores anotados referentes ao grau de similaridade semântica entre os pares de frases do dataset e os resultados obtidos a partir do modelo utilizado. Os resultados indicam uma importante consideração sobre a efetividade da engenharia de prompt na obtenção de melhores resultados com grandes modelos de linguagem.

Contrariando os achados desta pesquisa e corroborando com a literatura existente, observamos que o uso do Prompt Nível 01, embora tenha demonstrado melhor performance em termos de precisão, apresentou uma limitação significativa. Este prompt restringiu as respostas do modelo a uma faixa entre os valores 2,25 e 3,75. Esse comportamento sugere que o Prompt Nível 01 acabou "limitando" a fluidez e a capacidade do avançado modelo de inteligência artificial desenvolvido pela OpenAI, não explorando plenamente seu potencial.

Ao analisar os resultados do Prompt Nível 05, que incorpora orientações da engenharia de prompt e apresenta um nível de detalhamento mais completo, constata-se que o modelo conseguiu navegar por todo o espectro de respostas possíveis, obtendo valores que variam do mínimo (1,0) ao máximo (5,0). Isso demonstra que um prompt mais detalhado e bem elaborado permite que o modelo explore suas capacidades de maneira mais abrangente e eficaz.

Portanto, é crucial destacar que, mesmo com resultados divergentes nesta abordagem específica, a utilização da engenharia de prompt e o refinamento da redação do prompt podem proporcionar melhorias significativas nas respostas geradas pelo modelo. Este estudo ressalta que a aplicação de técnicas de engenharia de prompt pode favorecer uma utilização mais eficiente do potencial dos grandes modelos de linguagem, resultando em respostas mais precisas e alinhadas com as expectativas dos pesquisadores.

Além disso, a análise detalhada dos diferentes níveis de prompts revela que a

complexidade e a especificidade na redação dos prompts têm um impacto direto na qualidade das respostas. Prompts mais sofisticados, como o Nível 05, permitem que o modelo acesse uma gama mais ampla de respostas, refletindo uma maior capacidade de entendimento e processamento semântico. O fato enfatiza a importância de uma abordagem meticulosa na elaboração de prompts, que deve ser ajustada de acordo com as necessidades específicas da tarefa, para maximizar a eficiência e a precisão das respostas geradas.

Em suma, a investigação reforça a importância da engenharia de prompt na otimização do desempenho dos modelos de linguagem. Através do refinamento contínuo e da experimentação com diferentes níveis de detalhe e instrução, é possível desbloquear todo o potencial dos modelos de inteligência artificial, promovendo avanços significativos na área de Processamento de Linguagem Natural (PLN).

6 Considerações Finais

A área da inteligência artificial torna-se fundamental diante das necessidades que a modernidade impõe à sociedade em geral. O uso de IA é cada vez mais comum nas mais variadas áreas do conhecimento e é utilizada, muitas vezes, como aliada na execução de tarefas importantes do cotidiano. Uma das áreas fundamentais da inteligência artificial é o segmento de Processamento de Linguagem Natural, que tem como finalidade realizar análises referentes a dados textuais e de fala de forma eficiente. Diante da importância da comunicação, algo elementar humano, a utilização do Processamento de Linguagem Natural vem auxiliar nas lacunas na comunicação.

Com a chegada de Grandes Modelos de Linguagem, como o ChatGPT, há também a necessidade de formas eficazes de utilização para essas ferramentas tão poderosas. É nesse contexto que surge, de forma emergente, a Engenharia de Prompt, que orienta os usuários sobre como escrever os comandos inseridos em LLMs em busca de melhores respostas.

Dessa forma, esta pesquisa observou se a utilização de engenharia de prompt pode auxiliar positivamente nos resultados oriundos de Grandes Modelos de Linguagem, neste caso o ChatGPT, na atividade de comparação semântica entre frases na língua portuguesa brasileira. Para a análise, foram observados dois contextos.

O primeiro caso utilizou uma variedade de cinco prompts, construídos com auxílio da Engenharia de Prompt, visando a comparação semântica entre pares de frases. Nessa abordagem específica, observou-se que as diretrizes advindas da Engenharia de Prompt não colaboraram positivamente na obtenção de melhores respostas do modelo. Para chegar a essa observação, foram comparados os resultados dos diferentes prompts com os resultados anotados no dataset utilizado. Nesta avaliação, o Prompt 1, o mais simples e com comandos diretos, obteve maior proximidade dos resultados anotados. Os resultados foram analisados a partir das métricas Correlação de Pearson e Erro Quadrático Médio (MSE).

Na segunda abordagem, foram comparados os resultados dos prompts com os resultados de dez grupos de competidores do Workshop ASSIN II, cada grupo realizou três abordagens diferentes na tarefa de similaridade semântica. Obtivemos, neste âmbito, duas diferentes perspectivas. Ao compararmos a métrica Erro Quadrático Médio (MSE),

o Prompt 1, que atingiu o melhor resultado dentre as cinco abordagens no primeiro caso analisado, obteve a 2ª melhor colocação em comparação com os trinta resultados do ASSIN II. Já se tratando de Correlação de Pearson, o Prompt 1 obteve apenas a 23ª melhor colocação em comparação com os demais resultados do workshop.

Diante dos resultados, observa-se que a utilização de Engenharia de Prompt, no contexto de comparação de similaridade semântica, tende a melhorar os resultados obtidos. Mas, para isso, é fundamental a tarefa de refinamento dos prompts utilizados na execução. O refinamento dos prompts consiste na atividade de realizar mudanças na sua redação para que o modelo consiga gerar melhores respostas diante do que foi solicitado. Além disso, é necessário considerar o impacto da formulação do prompt nas predições do modelo, pois pequenas alterações podem levar a variações significativas nos resultados.

Sendo assim, destaca-se a necessidade de estudos aprofundados na Engenharia de Prompt e da realização da tarefa de testar modelos de prompts de acordo com a necessidade almejada. Um processo iterativo de ajuste e avaliação pode levar à criação de prompts mais eficazes, proporcionando respostas mais precisas e coerentes. Com isso, podemos não apenas melhorar a performance em tarefas específicas, como também contribuir para o avanço da área de Processamento de Linguagem Natural, promovendo a comunicação eficiente e precisa em contextos variados.

6.1 Limitações da pesquisa

O presente estudo, assim como é comum, apresentou algumas limitações em seu desenvolvimento. Tais limitações impactaram no prosseguimento da pesquisa, dificultando sua execução e adicionando mais complexidade ao processo.

O primeiro fator limitador foi que o tema analisado nesta pesquisa ainda está em viés muito inicial e torna-se ainda mais especifico quando se observa no contexto de análise semântica textual em língua portuguesa. Esse fato impactou negativamente devido aos poucos parâmetros existentes na literatura, no momento da realização da pesquisa, que poderiam ajudar em uma melhor escolha de tarefas durante a execução. A ausência de um maior conteúdo acadêmico sobre um estudo resulta em uma pesquisa muito inicial e suscetível a uma maior possibilidade de erros e problemas. Esta limitação se torna ainda mais grave ao se tratar do uso de Grandes Modelos de Linguagem para a finalidade de

similaridade semântica em língua portuguesa, uma área de estudo escassa e com poucos estudos realizados. A escassez de literatura sobre similaridade semântica em língua portuguesa brasileira resultou em dificuldades na escolha de tarefas mais apropriadas e na implementação de parâmetros adequados. Essa limitação pode ter levado a uma abordagem mais exploratória e inicial.

Além disso, a Engenharia de Prompt, ainda em seus estágios iniciais, contribuiu para a incerteza nos resultados obtidos. A falta de diretrizes específicas para a construção de prompts direcionados à similaridade semântica resultou em uma abordagem experimental, onde os mínimos detalhes nos prompts impactaram significativamente os resultados. A escassez de literatura sobre similaridade semântica em língua portuguesa brasileira deve-se ao fato de ser uma área de estudo emergente, com poucos estudos realizados até o momento.

Outra limitação da pesquisa desenvolvida foi a necessidade de refinamento dos prompts em busca de melhores resultados. A tarefa de melhoria da redação dos prompts é delicada e demorada, pois os mínimos detalhes podem impactar nos resultados gerados. Sendo assim, o tempo destinado à pesquisa foi um fator limitador, pois refinar prompts e testá-los diante da base de dados utilizada levava tempo e empregava custos financeiros para cada execução pela necessidade de utilização de uma API-key do OpenAI.

6.2 Trabalhos Futuros

Visando a superação das limitações identificadas, futuras pesquisas devem se concentrar na expansão do corpo de literatura sobre similaridade semântica em língua portuguesa brasileira. Investir em estudos mais aprofundados e diversificados permitirá o desenvolvimento de parâmetros mais robustos e a elaboração de melhores práticas. Além disso, é crucial continuar aprimorando a Engenharia de Prompt, no contexto da avaliação de similaridade semântica, desenvolvendo diretrizes mais específicas e testando diferentes abordagens para refinar os prompts, assim contribuindo com exemplos mais diretos para a área. Isso contribuirá para a obtenção de resultados mais precisos e confiáveis, permitindo uma melhor avaliação da eficácia dos Grandes Modelos de Linguagem em tarefas de similaridade semântica.

Referências

BARBOSA, R. de O.; TAVEIRA, F. A. L.; PERALTA, D. A. Entre respostas digitais e saberes experienciais: o chatgpt e a educação em perspectiva crítica. Revista Pesquisa Qualitativa, v. 12, n. 30, p. 01–18, 2024.

FENG, S.; YAN, X.; SUN, H.; FENG, Y.; LIU, H. X. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nature Communications, p. 14, 2021. Disponível em: <https://www.nature.com/articles/s41467-021-21007-8>. Acesso em: 27/12/2023.

FERRARO, D. S. e. S. B.; COELHO, M. A. Há como deter o ChatGPT? Uma resenha da obra de Lúcia Santaella. [S.l.]: SciELO Brasil, 2024.

FILHO, L. P.; SOUZA, T.; PAULA, L. Análise das respostas do chatgpt em relação ao conteúdo de programação para iniciantes. In: Anais do XXXIV Simpósio Brasileiro de Informática na Educação. Porto Alegre, RS, Brasil: SBC, 2023. p. 1738–1748. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/26794>. Acesso em: 10/07/2024.

FONSECA, E.; SANTOS, L.; CRISCUOLO, M.; ALUISIO, S. Assin: Avaliacao de similaridade semantica e inferencia textual. In: Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal. [s.n.], 2016. p. 13–15. Disponível em: <http://propor2016.di.fc.ul.pt/wp-content/uploads/2015/10/assin-overview.pdf>. Acesso em: 27/12/2023.

FONSECA, E. R.; SANTOS, L. B. dos; CRISCUOLO, M.; ALUÍSIO, S. M. Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática, v. 8, n. 2, p. 3–13, 2016. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/v8n2-1/377>. Acesso em: 07/08/2024.

FONSECA, J. J. S. da. Metodologia da pesquisa científica. Fortaleza: UEC, 2002. Apostila. Acesso em: 10/07/2024.

GIL, A. C. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, 2007.

KOJIMA, T.; GU, S. S.; REID, M.; MATSUO, Y.; IWASAWA, Y. Large language models are zero-shot reasoners. Advances in neural information processing systems, v. 35, p. 22199–22213, 2022. Disponível em: <https://arxiv.org/abs/2205.11916>. Acesso em: 20/11/2023.

LO, C. K. What is the impact of chatgpt on education? a rapid review of the literature. Education Sciences, v. 13, p. 410, 04 2023. Disponível em: <https://www.researchgate.net/publication/370090004_What_Is_the_Impact_of_ChatGPT_on_Education_A_Rapid_Review_of_the_Literature/citation/download>. Acesso em: 10/07/2024.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. Scielo, p. 11, 2021. Disponível em: <https://doi.org/10.1590/s0103-4014.2021.35101.007>. Acesso em: 27/12/2023.

NASCIMENTO, J. R. Exploração de técnicas de engenharia de prompt para aprimorar os resultados do uso de LLM no TCMRio. 2024. 60 p. Disponível em: <https://repositorio.ufrn.br/handle/123456789/58251>. Acesso em: 29/06/2024.

NAVEED, H.; KHAN, A. U.; QIU, S.; SAQIB, M.; ANWAR, S.; USMAN, M.; BARNES, N.; MIAN, A. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435, 2023. Disponível em: <https://arxiv.org/abs/2307.06435>. Acesso em: 13/07/2024.

NETO, M. P.; MIERS, C. Proposta de análise de desempenho do uso de mecanismos de segurança em sistemas inteligentes usando aprendizado de máquina e ia gerativa adversariais do tipo llm. In: Anais da XXIV Escola Regional de Alto Desempenho da Região Sul. Porto Alegre, RS, Brasil: SBC, 2024. p. 127–128. ISSN 2595-4164. Disponível em: <https://sol.sbc.org.br/index.php/eradrs/article/view/28026>. Acesso em: 10/06/2024.

OTTER, D. W.; MEDINA, J. R.; KALITA, J. K. A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems, v. 32, p. 604 – 624, 2020. Disponível em: <https://ieeexplore.ieee.org/document/9075398>. Acesso em: 27/12/2023.

PEREIRA, M. R.; BORDA, N. F.; MORALES, E. O. Inteligência artificial no cuidado: um desafio para a enfermagem. Enfermería: Cuidados Humanizados, Facultad de Ciencias de la Salud-Universidad Católica del Uruguay., v. 12, n. 1, 2023.

PINHO, C. M. d. A.; MOURA, A. F. d.; GASPAR, M. A.; NAPOLITANO, D. M. R. Identificação de deficiências em textos educacionais com a aplicação de processamento de linguagem natural e aprendizado de máquina. ETD Educação Temática Digital, UNICAMP, v. 24, n. 2, p. 350–372, 2022.

SANTU, S. K. K.; FENG, D. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. 2023. Disponível em: <https://arxiv.org/abs/2305.11430>. Acesso em: 21/11/2023.

SANT´ANA, F. P.; SANT´ANA, I. P.; SANT´ANA, C. d. C. Uma utilização do chat gpt no ensino. Com a Palavra, o Professor, v. 8, n. 20, p. 74–86, abr. 2023. Disponível em: <http://revista.geem.mat.br/index.php/CPP/article/view/951>.

SARTO, J. C.; QUADROS, S. F. P. et al. Integração da inteligência artificial na formação educacional médica: Oportunidades e desafios. 2024.

SHARMA, D.; KAUSHAL, S.; KUMAR, H.; GAINDER, S. Chatbots in healthcare: Challenges, technologies and applications. In: 4th International Conference on Artificial Intelligence and Speech Technology (AIST). [s.n.], 2022. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/10065328>. Acesso em: 10/07/2024.

SILVA, V.; FURTADO, E.; OLIVEIRA, J.; FURTADO, V. Engenharia de prompts em assistentes conversacionais para promoção de autocuidado baseados em modelos amplos de linguagem. In: Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde. Porto Alegre, RS, Brasil: SBC, 2024. p. 377–388. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/28833>. Acesso em: 10/07/2024.

SOUZA, F. A. de. Aplicação de Learning Analytics na Análise Automática de atividades de Introdução à Programação com o Scratch. 137 p. Dissertação (Mestrado) — Universidade Federal Rural de Pernambuco, Recife, 2022. Disponível em: <https://ww3.ppgia.ufrpe.br/sites/default/files/testes-dissertacoes/APLICA%C3%87%C3%83O%20DE%20LEARNING%20ANALYTICS%20NA%20AN%C3%81LISE.pdf>. Acesso em: 27/12/2023.

TAVARES, L. A.; MEIRA, M. C.; AMARAL, S. F. d. Inteligência artificial na educação: Survey / artificial intelligence in education: Survey. Brazilian Journal of Development, v. 6, n. 7, p. 48699–48714, jul. 2020. Disponível em: <https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/13539>.

TORRES, M. M. et al. Utilizando a llm gpt-3.5 turbo para o desenvolvimento de uma ferramenta de busca de materiais por características técnicas. Florianópolis, SC., 2023.

7 Apêndices

7.0.1 Evolução dos Prompts

Esta subseção demonstra alguns exemplos da evolução da construção dos prompts até o momento de geração do Prompt final que foi utilizado nos experimentos da pesquisa.

PROMPT - TÉCNICA ZERO-SHOT

NÍVEL 01:
Tabela de Referência de Similaridade:
1.0 - Pouco ou nada similar.
5.0 - Muito ou exatamente similar.
Frase A:
Frase B:
1. Compare as frases A e B e indique uma Pontuação de Similaridade entre elas seguindo a Tabela de Referência de Similaridade acima.

PROMPT - TÉCNICA FEW-SHOT

NÍVEL 01:
Exemplo 01 - Frases com Pontuação de Similaridade 1.0: Frase - Eu falo aqui no programa, e estou à disposição. Frase - Gilmar Rinaldi desafiou o senador Romário.
Exemplo 02 - Frases com Pontuação de Similaridade 2.75: Frase - A acusação foi apresentada pelo Ministério Público Federal (MPF) no início do mês. Frase - Todos passam a ser réus e começam a responder pelos crimes acusados pelo Ministério Público Federal (MPF).
Exemplo 03 - Frases com Pontuação de Similaridade 4.50: Frase - Fernanda Lima mostrou em sua conta no Instagram neste sábado, 12, seu novo corte de cabelo. Frase - Neste sábado (12) a apresentadora usou seu perfil no Instagram para exibir o corte de cabelo.

Tabela de Referência de Similaridade:

1.0 - Pouco ou nada similar.

5.0 - Muito ou exatamente similar.

Frase A:

Frase B:

1. Compare as frases A e B e indique uma Pontuação de Similaridade entre elas seguindo a Tabela de Referência de Similaridade acima.

2. Analise as respostas dos exemplos citados acima para dar sua resposta.

PROMPT - TÉCNICA ZERO-SHOT

(Utilizando o termo "Execute etapa por etapa")

NÍVEL 05:

Tabela de Referência de Similaridade:

1.0 - Pouco ou nada similar.

5.0 - Muito ou exatamente similar.

Frase A:

Frase B:

1. Compare as frases A e B e indique uma Pontuação de Similaridade entre elas seguindo a Tabela de Referência de Similaridade acima.

2. Observe a sintaxe e estrutura das palavras e aumente a pontuação de similaridade se as frases têm a mesma estrutura gramatical, ordem das palavras e construção das sentenças.

3. Observe a semântica e o significado e aumente a pontuação de similaridade se as frases expressam ideias similares ou idênticas, mesmo que as palavras sejam diferentes.

4.Indique palavras-chave e termos relevantes nas frases. Aumentando a pontuação de similaridade se essas características se repetem ou são semelhantes nas duas frases.

5. Analise o contexto e a intenção ao dar sua resposta, considerando aumentar a resposta de similaridade se as frases estão inseridas no mesmo contexto e a intenção comunicativa por trás delas denotam o mesmo sentido.

6. Analise os sinônimos das palavras contidas nas frases ao dar sua resposta.

7. Dê sua resposta de similaridade em um único valor, após considerar todos os itens acima.
8. Execute as tarefas etapa por etapa antes de dar sua resposta.

7.0.2 Prompts Utilizados nos Experimentos

Esta subseção demonstra os prompts utilizados no experimento final desta pesquisa.

PROMPT ZERO-SHOT - NÍVEL 01

Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.

1.0, 1.25, 1.50 ou 1.75: De ”As frases são completamente diferentes.”até ”As frases possuem pouca similaridade.”
2.0, 2.25, 2.50 ou 2.75: De ”As frases possuem pouca similaridade.”até ”As frases possuem similaridades.”
3.0, 3.25, 3.50 ou 3.75: De ”As frases possuem similaridades.”até ”As frases possuem muita similaridade.”
4.0, 4.25, 4.50 ou 4.75: De ”As frases possuem muita similaridade.”até ”As frases iguais ou são extremamente similares.”
5.0: As frases iguais ou são extremamente similares.

Frase A: frase1
Frase B: frase2

Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.

PROMPT ZERO-SHOT - NÍVEL 02

Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.

1.0, 1.25, 1.50 ou 1.75: De ”As frases são completamente diferentes.”até ”As frases possuem pouca similaridade.”
2.0, 2.25, 2.50 ou 2.75: De ”As frases possuem pouca similaridade.”até ”As frases possuem similaridades.”
3.0, 3.25, 3.50 ou 3.75: De ”As frases possuem similaridades.”até ”As frases possuem muita similaridade.”
4.0, 4.25, 4.50 ou 4.75: De ”As frases possuem muita similaridade.”até ”As frases iguais ou são extremamente similares.”
5.0: As frases iguais ou são extremamente similares.

Para gerar sua reposta, utilize as informações que estão entre triplos asteriscos abaixo:
***
Geralmente, frases com maior índice de similaridade possuem:
1º: Sintaxe e/ou a estrutura das palavras similares ou iguais.
***
Frase A: frase1
Frase B: frase2

Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.

PROMPT ZERO-SHOT - NÍVEL 03

Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.

1.0, 1.25, 1.50 ou 1.75: De ”As frases são completamente diferentes.”até ”As frases possuem pouca similaridade.”
2.0, 2.25, 2.50 ou 2.75: De ”As frases possuem pouca similaridade.”até ”As frases possuem similaridades.”
3.0, 3.25, 3.50 ou 3.75: De ”As frases possuem similaridades.”até ”As frases possuem muita similaridade.”
4.0, 4.25, 4.50 ou 4.75: De ”As frases possuem muita similaridade.”até ”As frases iguais ou são extremamente similares.”
5.0: As frases iguais ou são extremamente similares.

Para gerar sua reposta, utilize as informações que estão entre triplos asteriscos abaixo:
***
Geralmente, frases com maior índice de similaridade possuem:
1º: Sintaxe e/ou a estrutura das palavras similares ou iguais.
2º: Semântica e/ou o significado de cada palavra contidas nelas são similares ou iguais.
***
Frase A: frase1
Frase B: frase2

Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.

PROMPT ZERO-SHOT - NÍVEL 04

Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.

1.0, 1.25, 1.50 ou 1.75: De ”As frases são completamente diferentes.”até ”As frases possuem pouca similaridade.”
2.0, 2.25, 2.50 ou 2.75: De ”As frases possuem pouca similaridade.”até ”As frases possuem similaridades.”

3.0, 3.25, 3.50 ou 3.75: De ”As frases possuem similaridades.”até ”As frases possuem muita similaridade.”
4.0, 4.25, 4.50 ou 4.75: De ”As frases possuem muita similaridade.”até ”As frases iguais ou são extremamente similares.”
5.0: As frases iguais ou são extremamente similares.

Para gerar sua reposta, utilize as informações que estão entre triplos asteriscos abaixo:
***
Geralmente, frases com maior índice de similaridade possuem:
1º: Sintaxe e/ou a estrutura das palavras similares ou iguais.
2º: Semântica e/ou o significado de cada palavra contidas nelas são similares ou iguais.
3º: Crie palavras-chave e termos relevantes para cada frase. As frases possuem palavras-chave e/ou termos relevantes similares ou iguais.
***
Frase A: frase1
Frase B: frase2

Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.

PROMPT ZERO-SHOT - NÍVEL 05

Sua tarefa é indicar um valor de similaridade entre a Frase A e a Frase B. Use a pontuação abaixo para gerar sua resposta.

1.0, 1.25, 1.50 ou 1.75: De ”As frases são completamente diferentes.”até ”As frases possuem pouca similaridade.”
2.0, 2.25, 2.50 ou 2.75: De ”As frases possuem pouca similaridade.”até ”As frases possuem similaridades.”
3.0, 3.25, 3.50 ou 3.75: De ”As frases possuem similaridades.”até ”As frases possuem muita similaridade.”

4.0, 4.25, 4.50 ou 4.75: De ”As frases possuem muita similaridade.”até ”As frases iguais ou são extremamente similares.”
5.0: As frases iguais ou são extremamente similares.

Para gerar sua reposta, utilize as informações que estão entre triplos asteriscos abaixo:
***
Geralmente, frases com maior índice de similaridade possuem:
1º: Sintaxe e/ou a estrutura das palavras similares ou iguais.
2º: Semântica e/ou o significado de cada palavra contidas nelas são similares ou iguais.
3º: Crie palavras-chave e termos relevantes para cada frase. As frases possuem palavras-chave e/ou termos relevantes similares ou iguais.
4º: As frases possuem contexto e/ou intenção similares ou iguais;
5º: As frases possuem palavras que são sinônimos.
***
Frase A: frase1
Frase B: frase2

Sua resposta deverá conter apenas o valor numérico de similaridade entre cada par de frases, sem o uso de textos, podendo ser um número fracionado.