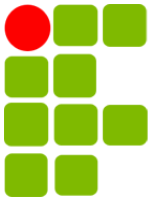


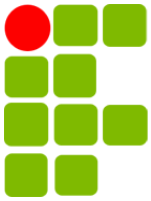
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Tecnologia em Telemática

Análise da Consistência de Modelos de Linguagem na Interpretação de Casos Clínicos

Juliana Miranda de Souto

Orientador: Ígor Barbosa da Costa

Campina Grande, Junho de 2026



Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Tecnologia em Telemática

Análise da Consistência de Modelos de Linguagem na Interpretação de Casos Clínicos

Juliana Miranda de Souto

Monografia apresentada à Coordenação do Curso de Telemática do IFPB - Campus Campina Grande, como requisito parcial para conclusão do curso de Tecnologia em Telemática.

Orientador: Ígor Barbosa da Costa

Campina Grande, Junho de 2026

Catalogação na fonte:
Ficha catalográfica elaborada por Thiago Ferreira Cabral de Oliveira – CRB 15/628

S728a SOUTO, Juliana Miranda de.
Análise da Consistência de Modelos de Linguagem na Interpretação de Casos Clínicos / Juliana Miranda de Souto. - Campina Grande, 2026.
38 f.

Monografia (Graduação em Tecnologia em Telemática) - Instituto Federal da Paraíba, 2026.
Orientador: Prof. Ígor Barbosa da Costa.

1. Saúde. 2. Inteligência Artificial. 3. Engenharia de Prompt. . 4. Processamento de Linguagem Natural. I. Costa, Ígor Barbosa da. II. Instituto Federal de Educação, Ciência e Tecnologia da Paraíba. III. Título.

CDU 004.9

Análise da Consistência de Modelos de Linguagem na Interpretação de Casos Clínicos

Juliana Miranda de Souto

Ígor Barbosa da Costa
Orientador

Mirna Carelli Oliveira Maia
Membro da Banca

Paulo Ribeiro Lins Júnior
Membro da Banca

Campina Grande, Paraíba, Brasil
Junho/2026

Aos meus pais e ao meu marido.
Sem vocês, nada disso seria possível.

"Só sei que nada sei."
– Sócrates (atribuído)

Agradecimentos

A conclusão deste trabalho é fruto de uma caminhada que começou muito antes da escrita destas páginas. Ela foi construída dia após dia, com o apoio de pessoas que tornaram cada obstáculo mais leve e cada conquista mais significativa.

Agradeço a todos os professores que cruzaram meu caminho desde o início da graduação. Cada um, com sua experiência e dedicação, deixou marcas que vão além do conteúdo das disciplinas. Em especial, ao meu orientador, Professor Ígor Barbosa da Costa, pela paciência, pelas valiosas orientações e por acreditar neste trabalho.

Aos meus pais,Teotônio Luiz de Souto e Maria José Miranda, minha eterna gratidão. Vocês são a minha base, o meu porto seguro. O incentivo constante, os sacrifícios silenciosos e o amor incondicional foram o alicerce que me permitiu chegar até aqui. Nada disso teria sentido sem vocês.

Ao meu marido, companheiro de todas as horas Humberto Antônio Santos da Silva, agradeço pela força que me dá em toda a caminhada – seja na vida acadêmica, seja na vida pessoal. Sua paciência, compreensão e incentivo foram essenciais nos momentos mais desafiadores.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, meu sincero obrigada.

Resumo

Este trabalho analisou a consistência de modelos de linguagem de grande escala na interpretação de casos clínicos. O objetivo foi investigar o comportamento dos modelos ChatGPT e DeepSeek ao responder repetidamente aos mesmos casos, utilizando diferentes estratégias de prompt, incluindo zero-shot e few-shot, em formatos estruturados e não estruturados. Para isso, foram utilizadas questões clínicas provenientes do dataset MedQA USMLE, selecionadas e filtradas com foco em diagnóstico. A consistência das respostas foi avaliada com base na estabilidade da hipótese principal, das hipóteses alternativas, do nível de urgência e da estrutura da resposta ao longo de múltiplas execuções. Os resultados indicaram que as estratégias de prompt influenciaram a estabilidade das respostas, sendo observadas maiores médias de consistência geral nas estratégias A e B. Também foi verificado que a hipótese principal e o nível de urgência apresentaram maior estabilidade do que as hipóteses alternativas. O estudo contribui para a compreensão da consistência como uma métrica complementar à acurácia na avaliação de modelos de linguagem em tarefas relacionadas à área da saúde.

Palavras-chave: Modelos de Linguagem, Consistência, Engenharia de Prompt, Inteligência Artificial em Saúde, Processamento de Linguagem Natural.

Abstract

This work analyzed the consistency of large language models in the interpretation of clinical cases. The objective was to investigate the behavior of ChatGPT and DeepSeek when repeatedly responding to the same cases using different prompting strategies, including zero-shot and few-shot, in both structured and unstructured formats. Clinical questions from the MedQA USMLE dataset were used, selected and filtered with a focus on diagnosis. Consistency was evaluated based on the stability of the main hypothesis, alternative hypotheses, urgency level, and response structure across multiple executions. The results indicated that prompting strategies influenced the stability of the generated responses, with strategies A and B presenting the highest overall consistency averages. It was also observed that the main hypothesis and urgency level were more stable than the alternative hypotheses. The study contributes to the understanding of consistency as a complementary metric to accuracy in the evaluation of language models in tasks related to healthcare.

Keywords: Language Models, Consistency, Prompt Engineering, Artificial Intelligence in Health, Natural Language Processing.

Sumário

Lista de Abreviaturas

IA	*Inteligência Artificial*
LLM	*Large Language Model* (Modelo de Linguagem de Grande Escala)
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
MedQA	*Medical Question Answering Dataset*
USMLE	*United States Medical Licensing Examination*

Lista de Tabelas

Capítulo 1

Introdução

O avanço da Inteligência Artificial (IA), especialmente no campo do Processamento de Linguagem Natural (Natural Language Processing – NLP), possibilita o desenvolvimento de modelos capazes de compreender e gerar linguagem natural com elevado nível de sofisticação. Entre esses modelos, destacam-se os Modelos de Linguagem de Grande Escala (Large Language Models – LLMs), baseados na arquitetura transformer proposta por [Vaswani *et al.* 2017], amplamente utilizados em tarefas que envolvem interpretação e geração de texto.

Nos últimos anos, esses modelos passaram a ser aplicados também na área da saúde, auxiliando na análise de registros clínicos, interpretação de sintomas e suporte à decisão médica. De acordo com [Shaik *et al.* 2023], o uso de IA em saúde tem se expandido significativamente, permitindo lidar com grandes volumes de dados e contribuir para a melhoria de processos clínicos.

Estudos recentes demonstram que modelos de linguagem podem apresentar desempenho relevante em tarefas clínicas, especialmente quando avaliados em bases de benchmark como o MedQA USMLE, amplamente utilizado na avaliação de sistemas de inteligência artificial em medicina [Singhal *et al.* 2024]. No entanto, apesar dos avanços em termos de desempenho, ainda existem desafios importantes relacionados ao comportamento desses modelos.

Uma dessas limitações refere-se à variabilidade das respostas. Trabalhos recentes indicam que modelos de linguagem podem apresentar variações significativas ao responder repetidamente ao mesmo problema, evidenciando a necessidade de avaliar não apenas a acurácia, mas também a consistência das respostas geradas [Azimi *et al.* 2025].

Apesar desses avanços, o uso de LLMs em tarefas relacionadas à área da saúde ainda apresenta limitações relevantes. Muitos estudos concentram-se na acurácia das respostas geradas, isto é, na capacidade do modelo de produzir respostas corretas. No entanto, em contextos clínicos, a avaliação da correção depende da análise de profissionais especializados, o que torna a acurácia um critério importante, mas não suficiente.

Nesse contexto, a consistência surge como uma métrica complementar. Diferentemente da acurácia, a consistência não busca determinar se uma resposta está correta ou incorreta, mas avaliar a estabilidade do comportamento do modelo ao longo de múltiplas execuções

para um mesmo caso clínico.

Por serem modelos probabilísticos, os LLMs podem gerar respostas diferentes para uma mesma entrada. Entretanto, variações acentuadas podem indicar instabilidade no comportamento do modelo, dificultando a compreensão de sua previsibilidade em cenários sensíveis, como tarefas relacionadas à área da saúde.

Dessa forma, a análise da consistência não tem como objetivo determinar a correção das respostas, mas sim avaliar o grau de variação entre elas, fornecendo uma medida de estabilidade do modelo.

Além disso, a forma como as instruções são fornecidas ao modelo, por meio da engenharia de prompt, pode influenciar diretamente o comportamento das respostas. Estudos demonstram que diferentes estratégias de prompt, como zero-shot, few-shot e prompts estruturados, podem impactar a qualidade, a organização e o comportamento das respostas geradas [Sonoda *et al.* 2024; Lopez *et al.* 2024; Sivarajkumar *et al.* 2024].

Apesar disso, ainda há necessidade de investigar de forma mais específica como essas estratégias afetam a consistência das respostas em tarefas clínicas.

Diante disso, este trabalho busca responder à seguinte questão de pesquisa: *como diferentes estratégias de engenharia de prompt influenciam a consistência das respostas geradas por modelos de linguagem na análise de casos clínicos?*

Para responder a essa questão, foi realizada uma análise experimental utilizando questões clínicas provenientes de um dataset público de benchmark. As questões foram submetidas a modelos de linguagem por meio de diferentes estratégias de prompt, com o objetivo de observar a estabilidade das respostas geradas em múltiplas execuções.

1.1 Justificativa e Relevância do Trabalho

A adoção de LLMs na área da saúde exige que esses sistemas apresentem não apenas respostas adequadas, mas também comportamento confiável e previsível. A consistência torna-se, portanto, um aspecto fundamental, especialmente em aplicações onde decisões podem impactar diretamente a vida humana.

Além disso, a relação entre consistência e acurácia não é direta. Um modelo pode apresentar respostas consistentes sem necessariamente estar correto, assim como pode produzir respostas corretas de forma instável. Por isso, consistência e acurácia devem ser compreendidas como dimensões distintas de avaliação.

Dessa forma, este trabalho contribui ao investigar a consistência como um critério independente, analisando como fatores externos, como a engenharia de prompt, influenciam o comportamento dos modelos.

1.2 Objetivos

1.2.1 Objetivo Geral

Analisar o impacto de diferentes estratégias de engenharia de prompt na consistência das respostas de modelos de linguagem aplicados à análise de casos clínicos.

1.2.2 Objetivos Específicos

- Quantificar a consistência das respostas geradas para um mesmo caso clínico;
- Avaliar o impacto de diferentes estratégias de prompt na estabilidade das respostas;
- Analisar indícios de sensibilidade dos modelos a diferentes estratégias de prompt;
- Identificar padrões de comportamento relacionados à variabilidade das respostas.

1.3 Metodologia

Este trabalho adota uma abordagem experimental baseada na análise de respostas geradas por modelos de linguagem a partir de casos clínicos.

Foram utilizadas diferentes estratégias de engenharia de prompt, incluindo abordagens zero-shot, few-shot e prompts estruturados. Para cada caso clínico, foram realizadas múltiplas execuções com o mesmo modelo e a mesma estratégia de prompt, permitindo avaliar a variabilidade das respostas.

A consistência foi medida com base na estabilidade das respostas geradas, considerando critérios semânticos e estruturais. Além disso, foram comparadas diferentes estratégias de prompt para analisar indícios de sensibilidade dos modelos às mudanças na forma de apresentação da entrada.

1.4 Organização do Documento

Este trabalho está organizado em cinco capítulos. O Capítulo 1 apresenta a introdução, a justificativa, os objetivos e uma visão geral da metodologia. O Capítulo 2 reúne os conceitos teóricos que fundamentam o estudo, abordando modelos de linguagem, aplicações na área da saúde, consistência e engenharia de prompt. O Capítulo 3 descreve a metodologia adotada, incluindo a base de dados, os modelos utilizados, as estratégias de prompt e as métricas de consistência. O Capítulo 4 apresenta os resultados obtidos e a análise comparativa entre modelos e estratégias de prompt. Por fim, o Capítulo 5 apresenta as conclusões do estudo e as possibilidades de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

2.1 Modelos de Linguagem de Grande Escala

Os modelos de linguagem de grande escala (Large Language Models – LLMs) são sistemas baseados em redes neurais treinados com grandes volumes de dados textuais, capazes de realizar tarefas como geração de texto, interpretação de linguagem natural e resposta a perguntas. Por operarem com mecanismos probabilísticos de geração, esses modelos podem produzir respostas diferentes para uma mesma entrada, especialmente quando submetidos a múltiplas execuções.

Essa característica não determinística está diretamente relacionada aos mecanismos internos de geração e representa um fator relevante na análise de consistência. Estudos mostram que estratégias de raciocínio e estruturação de entrada podem influenciar significativamente o comportamento desses modelos [Wang *et al.* 2024].

2.2 Aplicações de LLMs na Área da Saúde

Na área da saúde, os LLMs têm sido investigados em tarefas como interpretação de casos médicos, resposta a perguntas clínicas e apoio à análise de informações relacionadas ao cuidado em saúde. Avaliações recentes indicam que esses modelos apresentam desempenho promissor em benchmarks médicos, mas ainda enfrentam desafios relacionados à confiabilidade, previsibilidade e estabilidade de suas respostas [Singhal *et al.* 2024].

Além disso, o uso de inteligência artificial em tarefas relacionadas à saúde exige cuidado adicional, uma vez que interpretações inadequadas podem gerar riscos quando utilizadas sem avaliação profissional especializada [Shaik *et al.* 2023].

2.3 Consistência em Modelos de Linguagem

A consistência refere-se à capacidade de um modelo de produzir respostas semelhantes para uma mesma entrada ao longo de múltiplas execuções. Diferentemente da acurácia, que está

relacionada à correção da resposta, a consistência está associada à estabilidade do comportamento do modelo.

Estudos recentes demonstram que modelos de linguagem podem apresentar variações relevantes entre execuções, inclusive em tarefas clínicas, o que reforça a necessidade de avaliar essa propriedade de forma independente [Azimi *et al.* 2025].

Neste trabalho, a consistência é compreendida em duas dimensões principais: a consistência semântica, relacionada à estabilidade do conteúdo das respostas, como hipótese principal, hipóteses alternativas e nível de urgência; e a consistência estrutural, relacionada à manutenção do formato solicitado pelo prompt.

2.4 Engenharia de Prompt

A engenharia de prompt refere-se à elaboração de instruções utilizadas para orientar o comportamento dos modelos de linguagem. Estratégias como zero-shot e few-shot influenciam a forma como o modelo interpreta a entrada e organiza sua resposta [Lopez *et al.* 2024].

No contexto de tarefas clínicas, prompts mais estruturados podem contribuir para respostas mais organizadas e com campos mais bem definidos, como hipótese principal, hipóteses alternativas, nível de urgência e justificativa [Sonoda *et al.* 2024]. Dessa forma, a engenharia de prompt pode afetar não apenas a qualidade percebida da resposta, mas também sua estabilidade e padronização.

2.5 Relação entre Consistência e Engenharia de Prompt

Diferentes estratégias de prompt podem impactar não apenas o conteúdo das respostas, mas também sua estabilidade ao longo de múltiplas execuções. Mudanças na estrutura da instrução, na presença de exemplos ou na forma de solicitar a resposta podem levar a variações no comportamento dos modelos [Wang *et al.* 2024].

Dessa forma, analisar a relação entre engenharia de prompt e consistência permite compreender se determinadas formas de interação favorecem respostas mais estáveis em tarefas clínicas. Esse aspecto constitui o foco principal deste trabalho.

2.6 Desafios e Limitações

Apesar dos avanços, os LLMs apresentam limitações relevantes, especialmente no que diz respeito à consistência, confiabilidade e previsibilidade das respostas. A combinação de comportamento probabilístico e sensibilidade às instruções fornecidas pode dificultar a estabilidade das saídas geradas.

Esses desafios reforçam a necessidade de estudos experimentais que avaliem o comportamento dos modelos em diferentes condições de uso. No contexto deste trabalho, essa análise

é realizada a partir da comparação entre estratégias de prompt aplicadas a casos clínicos de um dataset público.

Capítulo 3

Metodologia

3.1 Visão Geral do Estudo

Este trabalho adota uma abordagem experimental com o objetivo de analisar a consistência das respostas geradas por modelos de linguagem de grande escala ao interpretar casos clínicos. A proposta consiste em submeter os modelos a diferentes estratégias de engenharia de prompt, aplicadas sobre os mesmos casos clínicos, e observar a variabilidade das respostas ao longo de múltiplas execuções.

Dessa forma, busca-se compreender como o comportamento dos modelos varia em função da forma de entrada fornecida, com foco na estabilidade das respostas, e não na avaliação de sua correção clínica.

3.2 Modelos Utilizados

Os experimentos foram realizados utilizando dois modelos de linguagem de grande escala: ChatGPT, disponibilizado pela OpenAI, e DeepSeek Chat, disponibilizado pela DeepSeek. Ambos foram acessados por meio de suas interfaces web no período de coleta dos dados.

A utilização de mais de um modelo permitiu comparar diferentes comportamentos diante dos mesmos estímulos, contribuindo para uma análise mais abrangente da consistência das respostas. Como os modelos foram utilizados por meio de interfaces web, os resultados refletem o comportamento observado nas versões disponíveis no período de realização dos experimentos.

3.3 Base de Dados

Os experimentos foram conduzidos utilizando questões provenientes do dataset MedQA USMLE, disponível publicamente.[1]

[1] <https://www.kaggle.com/datasets/evidence/medqa-usmle>

O dataset é composto por questões da área médica utilizadas em exames do tipo USMLE (United States Medical Licensing Examination), abrangendo diferentes especialidades clínicas.

Para este trabalho, foram selecionadas 50 questões, utilizadas como base para a construção dos casos analisados pelos modelos.

3.3.1 Seleção e preparação das questões

O dataset MedQA USMLE contém questões de diferentes áreas da medicina, incluindo diagnóstico, tratamento e prognóstico.

Para este trabalho, foram selecionadas apenas questões relacionadas a diagnóstico clínico, uma vez que esse tipo de questão está mais alinhado com o objetivo de analisar a consistência das respostas dos modelos.

A seleção das questões foi realizada por meio de scripts desenvolvidos em Python, utilizando o ambiente Google Colab. O processo consistiu na filtragem das questões com base em critérios específicos, como a presença de descrições clínicas e a necessidade de identificação de hipóteses diagnósticas.

Após a etapa de filtragem, foi definido um subconjunto de 50 questões, que foram utilizadas nos experimentos descritos neste trabalho.

As questões foram mantidas no idioma original, em inglês.

Neste estudo, foram utilizados apenas os enunciados clínicos das questões, sem as alternativas de múltipla escolha. Essa decisão foi tomada porque o objetivo não é avaliar a escolha da alternativa correta, mas a estabilidade da interpretação gerada pelo modelo.

O código utilizado para filtragem das questões encontra-se no Apêndice A.

3.4 Estratégias de Prompt

Foram utilizadas quatro estratégias de engenharia de prompt:

- **Prompt A – Zero-shot básico**: o modelo recebe apenas a questão clínica sem exemplos adicionais;

```
Você receberá um caso clínico para fins acadêmicos de pesquisa.
Analise o caso e informe:
1. hipótese principal
2. duas hipóteses alternativas
3. nível de urgência (baixa, média ou alta)
4. uma justificativa em até 4 linhas

Caso clínico: [CASO]
```

- **Prompt B – Zero-shot estruturado**: o modelo recebe a questão e é instruído a responder em formato estruturado;

```
Você receberá um caso clínico para fins acadêmicos de pesquisa.
Analise o caso e responda exatamente no formato abaixo:

Hipótese_principal:
Hipóteses_alternativas:
-
-
Nível_de_urgência:
Justificativa:

Caso clínico: [CASO]
```

- **Prompt C – Few-shot básico**: o modelo recebe exemplos prévios antes da questão;

```
Você receberá um caso clínico para fins acadêmicos de pesquisa.

Exemplo:

Caso clínico:
Paciente com febre, tosse e dor no corpo há 3 dias.

Resposta:
1. hipótese principal: infecção viral aguda
2. hipóteses alternativas: influenza; COVID-19
3. nível de urgência: média
4. justificativa: sintomas respiratórios agudos e febre recente

Agora analise:

Caso clínico: [CASO]
```

- **Prompt D – Few-shot estruturado**: o modelo recebe exemplos e também segue um formato estruturado de resposta.

```
Você receberá um caso clínico para fins acadêmicos de pesquisa.

Exemplo:
```

```
Caso clínico:
Paciente com febre, tosse e dor no corpo há 3 dias.

Resposta:
Hipótese_principal: infecção viral aguda
Hipóteses_alternativas:
- influenza
- COVID-19
Nível_de_urgência: média
Justificativa: sintomas respiratórios agudos e febre recente

Agora responda no mesmo formato:

Caso clínico: [CASO]
```

A classificação das estratégias foi utilizada apenas para fins de organização do experimento, não sendo explicitamente informada ao modelo durante a execução.

3.5 Configuração Experimental

Os experimentos foram realizados utilizando os modelos ChatGPT e DeepSeek, acessados por meio de suas interfaces web, no período de maio de 2026.

As execuções foram realizadas em sessões independentes, sem reutilização de histórico, de modo a evitar influência de interações anteriores nas respostas.

Não foi possível controlar diretamente parâmetros internos dos modelos, como temperatura, top-p e limite de tokens, devido às limitações das interfaces utilizadas. Essa restrição foi considerada na análise, uma vez que tais parâmetros podem influenciar a variabilidade das respostas.

As questões foram apresentadas no idioma original, em inglês, e os prompts foram escritos em português, mantendo-se essa configuração ao longo de todos os experimentos.

3.6 Procedimento de Execução

Para cada uma das 50 questões selecionadas, foram aplicadas quatro estratégias de prompt distintas: zero-shot básico, zero-shot estruturado, few-shot básico e few-shot estruturado.

Cada combinação de questão e tipo de prompt foi executada 3 vezes de forma independente, totalizando 150 execuções por tipo de prompt.

Considerando os quatro tipos de prompt, foram obtidas 600 respostas por modelo. Como o experimento foi conduzido utilizando dois modelos distintos, o total geral foi de 1200 respostas.

Em cada execução, o modelo foi solicitado a fornecer:

- Hipótese principal;
- Duas hipóteses alternativas;
- Nível de urgência;
- Justificativa.

Todas as respostas foram registradas para posterior análise de consistência.

3.7 Métricas de Consistência

Neste trabalho, a consistência foi definida como a capacidade do modelo de manter respostas semelhantes em diferentes execuções para o mesmo caso clínico.

A análise de consistência foi baseada na repetição dos elementos das respostas ao longo das execuções. Para cada critério avaliado, a consistência foi calculada pela frequência relativa da categoria mais recorrente entre as três execuções.

Dessa forma, quando as três execuções apresentaram a mesma categoria, a consistência foi igual a 1,00. Quando duas execuções apresentaram a mesma categoria, a consistência foi igual a 0,67.

Quando as três execuções apresentaram categorias distintas, a consistência foi igual a 0,33. A avaliação foi dividida em duas dimensões: consistência semântica e consistência estrutural.

3.7.1 Consistência Semântica

A consistência semântica refere-se à estabilidade do conteúdo das respostas geradas pelos modelos.

Foram avaliados os seguintes aspectos:

- Manutenção da hipótese principal ao longo das execuções;
- Similaridade entre as hipóteses alternativas;
- Estabilidade na classificação do nível de urgência.

A consistência semântica foi quantificada com base na frequência de repetição dos elementos ao longo das execuções.

Para a análise semântica, as respostas foram separadas por campo: hipótese principal, hipóteses alternativas, nível de urgência e justificativa. No caso das hipóteses diagnósticas, respostas com termos diferentes, mas significado equivalente, foram agrupadas manualmente em uma mesma categoria. Por exemplo, expressões que representavam a mesma hipótese clínica foram consideradas equivalentes, enquanto respostas com diferenças relevantes de interpretação foram mantidas em categorias distintas. O nível de urgência foi avaliado por correspondência direta entre as categorias baixa, média e alta. A consistência foi calculada pela frequência da categoria mais recorrente entre as execuções.

3.7.2 Consistência Estrutural

A consistência estrutural refere-se à capacidade do modelo de manter o formato de resposta solicitado nos prompts.

Foram considerados os seguintes critérios:

- Presença de todos os campos solicitados;
- Organização da resposta conforme o formato definido;
- Ausência de informações fora do padrão esperado.

Essa análise permitiu verificar se o modelo manteve não apenas o conteúdo, mas também a estrutura da resposta ao longo das execuções.

3.8 Análise dos Dados

Após a coleta das respostas geradas pelos modelos, foi realizada uma análise quantitativa da consistência, considerando diferentes dimensões.

Inicialmente, para cada questão, foram calculadas métricas de consistência com base nas múltiplas execuções realizadas para cada tipo de prompt. Em seguida, esses valores foram agregados para obter médias de consistência por:

- Tipo de prompt (zero-shot básico, zero-shot estruturado, few-shot básico e few-shot estruturado);
- Modelo utilizado (ChatGPT e DeepSeek);
- Dimensão de consistência (semântica e estrutural).

A análise também considerou a frequência de ocorrência dos elementos mais comuns nas respostas, como hipótese principal e nível de urgência, permitindo identificar padrões de estabilidade ou variação no comportamento dos modelos.

Além disso, foram realizadas comparações entre:

- Estratégias zero-shot e few-shot;
- Prompts básicos e estruturados;
- Diferentes modelos de linguagem.

Essas comparações tiveram como objetivo analisar quais configurações apresentaram maior estabilidade nas condições experimentais deste estudo.

Os resultados foram apresentados por meio de tabelas e medidas agregadas, facilitando a interpretação dos níveis de consistência observados.

3.9 Aspectos Éticos

Este trabalho utilizou exclusivamente dados públicos do dataset MedQA USMLE, disponível na plataforma Kaggle, que consiste em questões anônimas de exames médicos. Não foram utilizados dados reais de pacientes, prontuários ou informações pessoais identificáveis.

Os modelos de linguagem ChatGPT e DeepSeek foram utilizados exclusivamente como ferramentas de geração de respostas para os casos clínicos. A interação com os modelos se deu por meio de suas interfaces web padrão, sem acesso a parâmetros internos ou dados de treinamento.

As respostas geradas pelos modelos de linguagem foram analisadas apenas para fins acadêmicos e experimentais, não devendo ser interpretadas como recomendações clínicas ou substituição da avaliação por profissionais da área da saúde.

O trabalho respeita os princípios éticos relacionados ao uso responsável de inteligência artificial, especialmente no contexto da área médica.

Além disso, os LLMs foram empregados como ferramenta de apoio na formatação do texto, revisão gramatical e sugestão de estrutura. O conteúdo intelectual, a análise dos dados e a interpretação dos resultados são de inteira responsabilidade da autora.

Os princípios de transparência e reprodutibilidade foram respeitados: o código de filtragem das questões está disponível no Apêndice A, e as estratégias de prompt foram detalhadas na seção de metodologia.

3.10 Limitações do Método

Este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados.

Inicialmente, destaca-se o uso de um conjunto restrito de 50 questões, o que pode limitar a generalização dos resultados para outros contextos clínicos mais amplos.

Além disso, as questões são provenientes de exames padronizados (USMLE), não representando integralmente a complexidade de cenários clínicos reais.

Outra limitação relevante está relacionada à impossibilidade de controle de parâmetros internos dos modelos, como temperatura e top-p, o que pode ter influenciado a variabilidade das respostas.

Também deve ser considerado que os modelos utilizados podem sofrer atualizações ao longo do tempo, de modo que os resultados refletem o comportamento observado no período específico em que os experimentos foram realizados.

Por fim, não foi realizada validação clínica das respostas geradas, uma vez que a análise do trabalho é focada na consistência das respostas, e não em sua correção diagnóstica.

Capítulo 4

Resultados e Análise

Este capítulo apresenta os resultados obtidos a partir da análise das respostas geradas pelos modelos ChatGPT e DeepSeek para as 50 questões clínicas selecionadas. Para cada questão foram realizadas três execuções independentes em quatro estratégias distintas de engenharia de prompt (A, B, C e D), totalizando 1200 respostas analisadas (50 questões $\times$ 4 prompts $\times$ 3 execuções $\times$ 2 modelos).

Devido ao número reduzido de execuções por combinação experimental, os resultados devem ser interpretados como uma análise exploratória da consistência dos modelos nas condições avaliadas.

4.1 Procedimento de Cálculo da Consistência

Para cada combinação de questão, modelo e estratégia de prompt, foram comparadas as três respostas geradas.

A consistência foi calculada pela frequência relativa da categoria mais recorrente entre as três execuções. Dessa forma:

- Consistência = 1,00 quando as três execuções produziram a mesma categoria;
- Consistência = 0,67 quando duas execuções produziram a mesma categoria;
- Consistência = 0,33 quando as três execuções produziram categorias distintas.

Os valores apresentados nas tabelas correspondem à média dessas consistências ao longo das 50 questões avaliadas.

As respostas coletadas foram organizadas em uma planilha contendo modelo, questão, estratégia de prompt, execução, hipótese principal, hipóteses alternativas, nível de urgência, avaliação estrutural e categorias atribuídas durante a análise. Essa organização permitiu rastrear os cálculos apresentados nas tabelas.

4.2 Exemplo de Cálculo

A Tabela 4.1 apresenta um exemplo ilustrativo do procedimento utilizado.

Tabela 4.1: *Exemplo ilustrativo de cálculo da consistência*

Execução	Hipótese Principal	Urgência
1	Pneumonia bacteriana	Média
2	Pneumonia bacteriana	Média
3	Infecção viral respiratória	Baixa

Nesse exemplo, a consistência da hipótese principal foi 0,67, pois duas das três execuções apresentaram a mesma categoria. A consistência da urgência também foi 0,67 pelo mesmo motivo.

4.3 Visão Geral dos Resultados

A análise foi organizada em duas dimensões principais: consistência semântica e consistência estrutural.

A consistência semântica corresponde à média obtida a partir de três critérios: hipótese principal, hipóteses alternativas e nível de urgência. A consistência estrutural avalia a aderência das respostas ao formato esperado para cada estratégia de prompt.

Tabela 4.2: *Visão geral dos resultados consolidados*

Modelo	Consistência Semântica	Consistência Estrutural	Total de Respostas
ChatGPT	0,8467	0,6483	600
DeepSeek	0,7850	0,5700	600

Observa-se que ambos os modelos apresentaram níveis elevados de consistência semântica nas condições avaliadas. O ChatGPT apresentou médias superiores ao DeepSeek tanto na dimensão semântica quanto na estrutural. Esses resultados, entretanto, referem-se apenas à estabilidade das respostas, não à correção clínica das hipóteses geradas.

4.4 Consistência Semântica

A consistência semântica foi analisada separadamente para hipótese principal, hipóteses alternativas e classificação do nível de urgência.

4.4.1 Hipótese Principal

A hipótese principal apresentou os maiores níveis de consistência observados em todo o experimento. Os resultados sugerem elevada estabilidade na geração da hipótese diagnóstica

Tabela 4.3: *Consistência da Hipótese Principal por Modelo e Estratégia de Prompt*

Estratégia	ChatGPT	DeepSeek	Média
Prompt A	0,9933	0,9267	0,9600
Prompt B	0,9933	0,9400	0,9667
Prompt C	0,9800	0,8933	0,9367
Prompt D	0,9933	0,9467	0,9700
Média	**0,9900**	**0,9267**	**0,9583**

principal para ambos os modelos. Essa estabilidade indica repetição do comportamento observado entre execuções, mas não implica necessariamente correção diagnóstica.

4.4.2 Hipóteses Alternativas

Tabela 4.4: *Consistência das Hipóteses Alternativas por Modelo e Estratégia de Prompt*

Estratégia	ChatGPT	DeepSeek	Média
Prompt A	0,8267	0,6733	0,7500
Prompt B	0,8667	0,4533	0,6600
Prompt C	0,3333	0,3533	0,3433
Prompt D	0,3467	0,5133	0,4300
Média	**0,5933**	**0,4983**	**0,5458**

As hipóteses alternativas apresentaram níveis de consistência inferiores aos observados para a hipótese principal. Esse resultado sugere maior variabilidade dos modelos na geração de diagnósticos diferenciais, possivelmente porque diferentes hipóteses secundárias podem ser consideradas plausíveis para um mesmo caso clínico.

4.4.3 Nível de Urgência

Tabela 4.5: *Consistência do Nível de Urgência por Modelo e Estratégia de Prompt*

Estratégia	ChatGPT	DeepSeek	Média
Prompt A	0,9867	0,9400	0,9633
Prompt B	0,9333	0,9133	0,9233
Prompt C	0,9267	0,9067	0,9167
Prompt D	0,9800	0,9600	0,9700
Média	**0,9567**	**0,9300**	**0,9433**

A classificação do nível de urgência também apresentou elevada estabilidade, com médias superiores a 0,90 para ambos os modelos.

4.5 Consistência Estrutural

A consistência estrutural avaliou a capacidade dos modelos de manter o formato esperado para cada estratégia de prompt.

Tabela 4.6: *Consistência Estrutural por Modelo e Estratégia de Prompt*

Estratégia	ChatGPT	DeepSeek	Média
Prompt A	1,0000	1,0000	1,0000
Prompt B	0,9133	0,4600	0,6867
Prompt C	0,3333	0,3600	0,3467
Prompt D	0,3467	0,4600	0,4033
Média	**0,6483**	**0,5700**	**0,6092**

Observa-se maior estabilidade estrutural na estratégia A. Nas estratégias C e D houve redução da consistência estrutural para ambos os modelos.

4.6 Comparação entre Estratégias de Prompt

A consistência geral foi calculada como a média simples dos quatro indicadores analisados: consistência da hipótese principal, consistência das hipóteses alternativas, consistência do nível de urgência e consistência estrutural.

A Tabela 4.7 apresenta a média geral obtida por estratégia.

Tabela 4.7: *Consistência Geral por Modelo e Estratégia de Prompt*

Estratégia	ChatGPT	DeepSeek	Média
Prompt A	0,9517	0,8850	0,9183
Prompt B	0,9267	0,6917	0,8092
Prompt C	0,6433	0,6283	0,6358
Prompt D	0,6667	0,7200	0,6933

Os resultados sugerem que:

- A estratégia A apresentou a maior consistência geral observada no experimento;
- As estratégias A e B obtiveram médias superiores às estratégias C e D;
- O ChatGPT apresentou valores mais elevados nas estratégias A, B e C;
- O DeepSeek apresentou desempenho ligeiramente superior apenas na estratégia D.

4.7 Indícios de Sensibilidade às Estratégias de Prompt

A sensibilidade dos modelos às estratégias de prompt foi analisada a partir da variação das médias de consistência entre os prompts A, B, C e D. Observou-se que as estratégias produziram níveis diferentes de estabilidade, especialmente nas hipóteses alternativas e na consistência estrutural.

De modo geral, os resultados indicam que a forma de apresentação da entrada influenciou o comportamento dos modelos. No entanto, essa análise deve ser interpretada como exploratória, uma vez que o estudo não teve como objetivo isolar todos os fatores que podem afetar a variabilidade das respostas, como parâmetros internos de geração, atualizações dos modelos ou diferenças entre interfaces web.

4.8 Discussão dos Resultados

Os resultados obtidos sugerem que ambos os modelos apresentaram elevada estabilidade na geração da hipótese principal e na classificação do nível de urgência, considerando as condições experimentais adotadas neste estudo.

A menor consistência observada nas hipóteses alternativas indica que os diagnósticos diferenciais tendem a apresentar maior variabilidade entre execuções. Esse comportamento é esperado, uma vez que diferentes possibilidades diagnósticas podem ser consideradas plausíveis para um mesmo caso clínico.

Também foi observada variação entre as estratégias de prompt avaliadas. Nas condições deste experimento, as estratégias A e B apresentaram maiores níveis de consistência geral quando comparadas às estratégias C e D. Esse resultado indica que, neste conjunto de questões, a inclusão de exemplos nos prompts few-shot não levou necessariamente a maior estabilidade das respostas.

Quanto à comparação entre modelos, o ChatGPT apresentou maior consistência em grande parte das métricas avaliadas. Entretanto, como os modelos foram acessados por interfaces web e não houve controle sobre parâmetros internos de inferência, não é possível determinar as causas dessas diferenças observadas. Dessa forma, os resultados devem ser interpretados como observações empíricas restritas às condições deste estudo.

Por fim, os resultados apresentados indicam que a engenharia de prompts pode influenciar a estabilidade das respostas geradas por modelos de linguagem em tarefas clínicas. Assim, a definição da estratégia de interação utilizada deve ser considerada um fator relevante em estudos que avaliam o comportamento desses modelos, especialmente quando o foco está na previsibilidade e na consistência das respostas.

4.9 Escopo da Análise e Ameaças à Validade

Os resultados apresentados neste capítulo devem ser interpretados considerando o caráter exploratório do experimento. O objetivo não foi estabelecer conclusões definitivas sobre qual modelo ou estratégia de prompt é mais consistente, mas observar padrões de estabilidade e variabilidade nas condições específicas avaliadas.

Embora tenham sido analisadas 1200 respostas, essas respostas não correspondem a 1200 observações independentes, pois estão organizadas em torno de 50 questões clínicas, quatro estratégias de prompt, dois modelos e três execuções por combinação experimental. Dessa forma, as três execuções representam repetições de uma mesma condição, enquanto as questões clínicas constituem a principal unidade de análise. O número reduzido de execuções limita a capacidade de caracterizar de forma abrangente a variabilidade dos modelos.

Outra limitação decorre do uso de interfaces web para acesso aos modelos. Não houve controle direto sobre parâmetros internos de geração, como temperatura, mecanismos de amostragem, mensagens de sistema ou versão específica do modelo utilizada em cada execução. Também não é possível descartar completamente a ocorrência de atualizações dos sistemas durante o período de coleta. Consequentemente, diferenças observadas entre modelos ou estratégias de prompt não podem ser atribuídas exclusivamente às características dos prompts avaliados.

A operacionalização da consistência também constitui uma ameaça à validade de construto. As respostas textuais foram agrupadas em categorias para permitir sua comparação, o que exige decisões relacionadas à equivalência entre termos, sinônimos, níveis de especificidade e hipóteses clinicamente semelhantes. Essa questão é particularmente relevante para as hipóteses alternativas, que podem variar em quantidade, ordem e nível de detalhamento. Assim, parte da variabilidade observada pode estar relacionada tanto ao comportamento dos modelos quanto ao procedimento de categorização adotado.

A medida utilizada resume a frequência da categoria predominante entre três execuções. Essa operacionalização permite comparar as condições avaliadas, mas representa uma simplificação da concordância entre as respostas. Portanto, seus valores devem ser compreendidos como indicadores de estabilidade relativa dentro do protocolo utilizado, e não como medidas universais de consistência dos modelos.

Também deve ser considerada a distinção entre consistência e correção clínica. Uma resposta pode ser reproduzida de maneira estável e, ainda assim, apresentar uma hipótese inadequada. Como o experimento não comparou sistematicamente as respostas com diagnósticos de referência ou avaliações independentes de especialistas, os resultados não permitem inferir acurácia diagnóstica, segurança clínica ou qualidade das recomendações produzidas.

Por fim, os resultados estão limitados aos 50 casos clínicos, aos quatro prompts, aos dois modelos, às interfaces e ao período de coleta considerados. Não se pode assumir que os mesmos padrões seriam observados em outros modelos, versões, especialidades médicas, conjuntos de questões ou ambientes de execução.

Diante dessas limitações, as diferenças identificadas devem ser interpretadas como indícios experimentais que podem orientar estudos posteriores. Experimentos futuros poderão ampliar o número de execuções, utilizar acesso por API com parâmetros controlados, registrar versões específicas dos modelos, empregar avaliadores independentes e incluir medidas de correção clínica.

Capítulo 5

Conclusão

5.1 Conclusões do Estudo

Este trabalho analisou como diferentes estratégias de engenharia de prompt influenciam a consistência das respostas de modelos de linguagem na interpretação de casos clínicos. O foco da análise não foi avaliar a correção clínica das respostas geradas, mas sim observar a estabilidade do comportamento dos modelos ao longo de múltiplas execuções.

Os resultados indicaram que a forma como o prompt é elaborado pode influenciar a consistência das respostas. Nas condições experimentais avaliadas, as estratégias A e B apresentaram as maiores médias de consistência geral. Também foi observado que a hipótese principal e a classificação do nível de urgência tenderam a apresentar maior estabilidade do que as hipóteses alternativas.

A comparação entre os modelos ChatGPT e DeepSeek mostrou diferenças de comportamento diante dos mesmos casos e estratégias de prompt. De modo geral, o ChatGPT apresentou médias superiores na maior parte das métricas analisadas. No entanto, esses resultados devem ser interpretados com cautela, pois os modelos foram acessados por meio de interfaces web, sem controle direto de parâmetros internos de geração, como temperatura e top-p.

Os achados reforçam que a consistência é uma dimensão relevante na avaliação de modelos de linguagem, especialmente em tarefas relacionadas à área da saúde. Um modelo pode apresentar respostas estáveis sem necessariamente estar clinicamente correto, assim como pode gerar respostas corretas de forma instável. Por isso, a consistência deve ser compreendida como uma métrica complementar à acurácia.

Como contribuição, este trabalho apresenta uma análise experimental da estabilidade de respostas de LLMs em casos clínicos, comparando diferentes estratégias de prompt e dois modelos distintos. Além disso, o código de filtragem das questões está disponível no Apêndice A, e as estratégias de prompt foram descritas em detalhe, favorecendo a replicação do estudo por outros pesquisadores.

5.2 Trabalhos Futuros

Com base nas limitações identificadas, sugerem-se as seguintes possibilidades para continuidade da pesquisa:

1. **Controle de parâmetros internos**: repetir os experimentos utilizando APIs que permitam configurar temperatura, top-p, limite de tokens e outros parâmetros de inferência.

2. **Ampliação da base de casos**: expandir o número de questões analisadas, incluindo maior diversidade de especialidades médicas e níveis de complexidade.

3. **Inclusão de outros modelos**: avaliar modelos adicionais, incluindo LLMs generalistas e modelos especializados em tarefas biomédicas ou clínicas.

4. **Correlação entre consistência e acurácia**: investigar se respostas mais consistentes também tendem a ser clinicamente corretas, com apoio de profissionais da área da saúde.

5. **Avaliação com maior número de execuções**: aumentar o número de repetições por caso clínico e por estratégia de prompt, permitindo uma análise mais robusta da variabilidade das respostas.

6. **Sensibilidade a pequenas variações no prompt**: investigar como mudanças mínimas na redação das instruções afetam a estabilidade das respostas geradas.

7. **Aplicação a outras tarefas clínicas**: estender o estudo para tarefas como sumarização de informações clínicas, interpretação de exames ou recomendação de condutas, sempre considerando a necessidade de validação especializada.

Apêndice A

Código de Pré-processamento dos Dados

O código a seguir foi utilizado para realizar o download e filtragem das questões do dataset MedQA, selecionando apenas aquelas relacionadas a diagnóstico clínico.

```
!pip install gdown

import gdown
import json
import zipfile
import os

# Fonte: https://github.com/jind11/MedQA
google_drive_url = "https://drive.google.com/file/d/1ImYUSLk9JbgHXOemfvyiDiirluZHPeQw/view?usp=sharing"

# ID do arquivo
file_id = "1ImYUSLk9JbgHXOemfvyiDiirluZHPeQw"
download_url = f"https://drive.google.com/uc?id={file_id}"

print("Baixando dataset do Google Drive (link oficial)...")
print(f"URL: {download_url}")
print("Isso pode levar alguns minutos (arquivo ~314 MB)...")

#  baixar o arquivo
!gdown {download_url}

# nome do arquivo baixado
print("\nArquivos baixados:")
```

```
import json
import pandas as pd

palavras_diagnostico = [
    "most likely diagnosis",
    "most likely cause",
    "etiology",
```

```
    "most consistent with"
]

palavras_excluir = [
    "treatment", "therapy", "medication",
    "test", "mri", "ct scan"
]

def eh_apenas_diagnostico(questao):
    pergunta = questao.get('question', '').lower()
    tem_diag = any(p in pergunta for p in palavras_diagnostico)
    tem_excl = any(p in pergunta for p in palavras_excluir)
    return tem_diag and not tem_excl

todas_questoes = []
with open('train.jsonl', 'r', encoding='utf-8') as f:
    for linha in f:
        if linha.strip():
            todas_questoes.append(json.loads(linha))

questoes_diagnostico = [q for q in todas_questoes if eh_apenas_diagnostico
    (q)]

selecionadas = questoes_diagnostico[:400]

df = pd.DataFrame(selecionadas)
df.to_csv('medqa_filtrado.csv', index=False)
```

Listing A.1: *Filtragem de questões do MedQA*

Referências Bibliográficas

[Azimi *et al.* 2025] AZIMI, I. *et al.* Evaluation of large language models accuracy and consistency in medical question answering. *Scientific Reports*, 2025. 1, 5

[Lopez *et al.* 2024] LOPEZ, M. *et al.* Prompt engineering principles for generative ai use. *Journal of Education and Learning*, 2024. 2, 5

[Shaik *et al.* 2023] SHAIK, T. *et al.* Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2023. 1, 4

[Singhal *et al.* 2024] SINGHAL, K. *et al.* Benchmarking large language models in evidence-based medicine. *IEEE Journal of Biomedical and Health Informatics*, 2024. 1, 4

[Sivarajkumar *et al.* 2024] SIVARAJKUMAR, S. *et al.* An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 2024. 2

[Sonoda *et al.* 2024] SONODA, T. *et al.* Structured clinical reasoning prompt enhances large language model's diagnostic capabilities. *Internal Medicine*, 2024. 2, 5

[Vaswani *et al.* 2017] VASWANI, A. *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2017. 1

[Wang *et al.* 2024] WANG, X. *et al.* A comparison of chain-of-thought reasoning strategies in large language models. *PeerJ Computer Science*, 2024. 4, 5

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande - Código INEP: 25137409
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

TCC

Assunto:	TCC
Assinado por:	Juliana Miranda
Tipo do Documento:	Anexo
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Juliana Miranda de Souto**, **DISCENTE (202121210027) DE TECNOLOGIA EM TELEMÁTICA - CAMPINA GRANDE**, em 03/09/2026 13:48:53.

Este documento foi armazenado no SUAP em 03/09/2026. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1976462
Código de Autenticação: 0d4e8abc19