



UNIVERSIDADE FEDERAL DO CEARÁ
INSTITUTO UNIVERSIDADE VIRTUAL
CURSO DE GRADUAÇÃO EM SISTEMAS E MÍDIAS DIGITAIS

YANNA TORRES GONÇALVES

**DESENVOLVIMENTO DE UM LARGE LANGUAGE MODEL: UMA ESTRATÉGIA
DE APRIMORAMENTO DE TRANSCRIÇÕES DE ANAMNESES MÉDICAS**

FORTALEZA

2024

YANNA TORRES GONÇALVES

DESENVOLVIMENTO DE UM LARGE LANGUAGE MODEL: UMA ESTRATÉGIA DE
APRIMORAMENTO DE TRANSCRIÇÕES DE ANAMNESES MÉDICAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Orientador: Prof. Dr. Rafael Augusto Ferreira do Carmo.

Coorientadora: Prof^a. Dra. Ticiano Linhares Coelho da Silva.

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

G629d Gonçalves, Yanna Torres.

Desenvolvimento de um large language model : uma estratégia de aprimoramento de transcrições de anamneses médicas / Yanna Torres Gonçalves. – 2024.
52 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Instituto UFC Virtual, Curso de Sistemas e Mídias Digitais, Fortaleza, 2024.

Orientação: Prof. Dr. Rafael Augusto Ferreira do Carmo.

Coorientação: Profa. Dra. Ticiania Linhares Coelho da Silva.

1. Modelos grandes de linguagem (LLMs). 2. Anamnese médica. 3. Text style transfer (TST). 4. Aprendizado de máquina. I. Título.

CDD 302.23

YANNA TORRES GONÇALVES

DESENVOLVIMENTO DE UM LARGE LANGUAGE MODEL: UMA ESTRATÉGIA DE
APRIMORAMENTO DE TRANSCRIÇÕES DE ANAMNESES MÉDICAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Aprovada em: 31/01/2025.

BANCA EXAMINADORA

Prof. Dr. Rafael Augusto Ferreira do
Carmo (Orientador)
Universidade Federal do Ceará (UFC)

Prof^a. Dra. Ticianá Linhares Coelho da
Silva (Coorientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. José Antônio Fernandes de Macêdo
Universidade Federal do Ceará (UFC)

Para o meu eu de ontem, que sonhou; para o meu eu de hoje, que persevera; e para o meu eu de amanhã, que ainda alcançará grandes conquistas.

AGRADECIMENTOS

À professora Ticiania pela constante inspiração e pelo incentivo contínuo em tudo.

Ao Professor José Macêdo pelas oportunidades na área de pesquisa.

Ao João Victor e ao Breno pela parceria e apoio durante o projeto CEREIA, assim como à Maria Alice, à Ana Alice e ao Lucas.

Aos colegas de curso, por todas as risadas compartilhadas e pelo apoio mútuo em momentos desafiadores.

Aos professores com quem tive a oportunidade de ter aulas ou trabalhar com, agradeço pelo conhecimento e pelas valiosas lições transmitidas.

Aos amigos de infância, Filipe e Jamilly, pelo apoio incondicional ao longo de toda a trajetória.

À minha irmã e ao meu irmão, pelo apoio em todos os momentos.

Às minhas sobrinhas, Helena e Júlia, que sempre são uma fonte de alegria.

Aos meus pais, cujos sacrifícios e dedicação tornaram possível a realização deste sonho.

Por fim, agradeço a todos que, de alguma maneira, contribuíram para a realização deste trabalho e conclusão do curso.

"And though I can't recall your face
I still got love for you"
(Taylor Swift)

RESUMO

Este trabalho tem como objetivo explorar o desenvolvimento de um Large Language Model (LLM) para aprimorar transcrições de anamneses médicas. A pesquisa envolveu duas etapas principais: a primeira consistiu em um *benchmark* inicial de três modelos de linguagem (Phi3, Llama e Mistral), avaliando seu desempenho na tarefa de formatação de anamneses, e a segunda focou na aplicação de *fine-tuning* no modelo Mistral, com base em críticas qualitativas e quantitativas dos avaliadores. O modelo Mistral se destacou no *benchmark* devido à sua consistência e melhores resultados nas métricas de concordância, como os coeficientes de Kendall e Kappa, embora tenha sido criticado pela sobrecarga de detalhes em suas respostas, prejudicando sua objetividade. O modelo Llama apresentou o pior desempenho, com respostas frequentemente irrelevantes, enquanto o Phi3 teve um desempenho moderado, mas com falhas em precisão e uso excessivo de linguagem genérica. O *fine-tuning* aplicado ao modelo Mistral, apesar de tentar corrigir as falhas, não resultou em melhorias significativas, já que a média das avaliações foi ligeiramente inferior à do modelo original. As principais contribuições deste trabalho incluem a comparação de modelos de linguagem para uma tarefa específica da área médica e a análise das limitações do processo de *fine-tuning*. O estudo também identificou limitações importantes, como a qualidade dos dados e o viés potencial entre os avaliadores, que podem ter influenciado os resultados. Como trabalhos futuros, pretende-se realizar treinamentos mais longos, expandir e diversificar o conjunto de dados e explorar novas abordagens de aprendizado de máquina para melhorar ainda mais o desempenho do modelo.

Palavras-chave: Modelos Grandes de Linguagem (LLMs); Anamnese Médica; Text Style Transfer (TST); Aprendizado de Máquina.

ABSTRACT

This work aims to explore the development of a Large Language Model (LLM) to enhance the transcription of patient records. The research involved two main stages: the first consisted of an initial benchmark of three language models (Phi3, Llama, and Mistral), evaluating their performance in the task of formatting anamneses, and the second focused on fine-tuning the Mistral model, based on qualitative and quantitative feedback from the evaluators. The Mistral model stood out in the benchmark due to its consistency and better results in agreement metrics, such as Kendall's tau and Fleiss' Kappa coefficients, although it was criticized for overloading details in its responses, impairing its objectivity. The Llama model performed the worst, with responses that were often irrelevant, while Phi3 showed moderate performance, but with issues in precision and excessive use of generic language. The fine-tuning of the Mistral model, despite attempts to correct the flaws, did not result in significant improvements, as the average evaluation score was slightly lower than the original model. The main contributions of this work include the comparison of language models for a specific task in the medical field and the analysis of the limitations of the fine-tuning process. The study also identified important limitations, such as data quality and potential bias among the evaluators, which may have influenced the results. As future work, the intention is to conduct longer training sessions, expand and diversify the dataset, and explore new machine learning approaches to further improve the model's performance.

Keywords: Large Language Models (LLMs); Patient Records; Text Style Transfer (TST); Machine Learning.

LISTA DE FIGURAS

Figura 1 – Estrutura básica de um ASR	20
Figura 2 – Abordagem do Wav2Vec 2.0	21
Figura 3 – Abordagem do Whisper	22
Figura 4 – Similaridade Semântica entre 'chocolate' e outras palavras	23
Figura 5 – Visão Geral do Fluxo de Etapas de um LLM	25
Figura 6 – LLMs lançados a partir de 2019 em ordem cronológica: modelos em azul representam 'pré-treinados', enquanto modelos em laranja são 'instruction-tuned'	27
Figura 7 – Diagrama de Venn da definição linguística de estilo e definição de estilo baseada em dados.	27
Figura 8 – Arquitetura e treinamento do Llama 3	29
Figura 9 – Processo de Desenvolvimento	31
Figura 10 – Prompt padrão implementado no modelo phi3	32
Figura 11 – Prompt utilizado para o treinamento	34
Figura 12 – Exemplo de transcrição utilizada durante o processo de avaliação	38
Figura 13 – Exemplo de formatação feita pelo Phi3	40
Figura 14 – Exemplo de formatação feita pelo Llama	42
Figura 15 – Exemplo de formatação feita pelo Mistral	43
Figura 16 – Exemplo de formatação feita pelo Mistral com Fine-tuning	45

LISTA DE TABELAS

Tabela 1 – Hiperparâmetros do treinamento.	35
Tabela 2 – Resultados descritivos do modelo Phi3	39
Tabela 3 – Coeficiente de correlação <i>tau</i> de Kendall para o modelo Phi3	39
Tabela 4 – Resultados descritivos do modelo Llama	40
Tabela 5 – Coeficiente de correlação <i>tau</i> de Kendall para o modelo Llama	41
Tabela 6 – Resultados descritivos do modelo Mistral	41
Tabela 7 – Coeficiente de correlação <i>tau</i> de Kendall para o modelo Mistral	41
Tabela 8 – Resultados Descritivos do modelo Mistral com Fine-tuning	44
Tabela 9 – Coeficiente de Correlação tau de Kendall entre avaliadores para o modelo Mistral com Fine-tuning	44

LISTA DE ABREVIATURAS E SIGLAS

ASR	<i>Automatic Speech Recognition</i>
CoT	<i>Chain-of-Thought</i>
GQA	<i>Grouped-Query Attention</i>
HDA	História da Doença Atual
HF	História Familiar
HPP	História Patológica Pgressa
HPS	História Pessoal e Social
IoT	Internet das Coisas
LLM	<i>Large Language Model</i>
LM	<i>Language Model</i>
MoE	<i>Mixture-of-Experts</i>
PLN	Processamento de Linguagem Natural
QP	Queixa Principal
RS	Revisão de Sistemas
SVA	<i>Smart Voice Assistants</i>
SWA	<i>Sliding Window Attention</i>
TST	Transferência de Estilo Textual
WER	<i>Word Error Rate</i>

SUMÁRIO

1	INTRODUÇÃO	12
2	REFERENCIAL TEÓRICO	15
2.1	Anamnese Médica	15
2.1.1	<i>Estrutura Base das Anamneses</i>	16
2.1.2	<i>Particularidades do Texto e Desafios</i>	18
2.2	Automatic Speech Recognition	19
2.2.1	<i>Modelos do Estado da Arte</i>	20
2.2.2	<i>Desafios</i>	22
2.3	Modelos Grandes de Linguagem	23
2.3.1	<i>Prompt Engineering</i>	25
2.3.2	<i>Instruct Based Models</i>	26
2.3.3	<i>Transferência de Estilo Textual</i>	26
2.3.4	<i>Modelos do Estado da Arte</i>	28
3	METODOLOGIA	31
3.1	Benchmark	31
3.1.1	<i>Elaboração do Prompt</i>	31
3.1.2	<i>Modelos Avaliados</i>	32
3.2	Coleta e Geração de Dados de Treinamento	33
3.3	Fine-tuning	34
3.4	Avaliação	35
3.4.1	<i>Coefficiente de Correlação tau de Kendall</i>	36
3.4.2	<i>Kappa de Fleiss</i>	37
4	RESULTADOS	38
4.1	Resultados do <i>Benchmark</i>	38
4.1.1	<i>Modelo Phi3</i>	38
4.1.2	<i>Modelo Llama</i>	39
4.1.3	<i>Modelo Mistral</i>	41
4.2	Resultados do <i>Fine-tuning</i>	43
5	CONCLUSÕES E TRABALHOS FUTUROS	46
	REFERÊNCIAS	48

1 INTRODUÇÃO

Em meio às práticas médicas contemporâneas, a anamnese surge como um elemento crucial na avaliação e diagnóstico dos pacientes. De acordo com Soares *et al.* (2016), a construção de uma anamnese detalhada é uma etapa fundamental no processo de avaliação do paciente, orientando a conduta clínica e a tomada de decisões médicas. Esse processo permite ao profissional de saúde obter uma visão abrangente do paciente, considerando desde aspectos físicos até os sociais e psicológicos.

Entretanto, o cenário atual dentro dos consultórios e hospitais apresenta desafios significativos. Conforme relatado pela Hapvida NotreDame Intermédica¹, aproximadamente 50% do tempo dedicado a uma consulta é consumido pela documentação da anamnese. Adicionalmente, as anamneses frequentemente carecem de detalhamento e abrangência, o que pode resultar em diagnósticos imprecisos e tratamentos inadequados.

Esse cenário é amplamente influenciado pelos processos burocráticos envolvidos na coleta de informações. Além disso, durante a fase de documentação da anamnese, os profissionais de saúde podem enfrentar dificuldades para recordar todos os detalhes relevantes fornecidos pelo paciente, podendo omitir detalhes que podem ser cruciais mais tarde, o que influencia diretamente na qualidade final da anamnese (Gür, 2012).

Diante desses desafios, uma solução promissora para otimizar esse processo é a adoção de modelos de reconhecimento automático de fala, ou *Automatic Speech Recognition* (ASR). ASR é a tecnologia que permite que um computador reconheça e interprete a fala humana, transformando-a em texto (Reddy, 1976). Atualmente, um dos principais usos dessa tecnologia são os Assistentes de Voz Inteligentes, ou *Smart Voice Assistants* (SVA), que podem ser encontrados em uma variedade de dispositivos, como *smartphones*, sistemas de comando de voz em carros e em dispositivos de Internet das Coisas (IoT), além de softwares, como motores de busca, legendas automáticas e sistemas de resposta de voz interativa (Ngueajio; Washington, 2022).

No campo médico, sistemas que utilizam ASR têm sido alvo de testes e estudos em uma variedade de cenários. Alguns exemplos de uso são na melhoria das habilidades de leitura em crianças com Síndrome de Down (Felix *et al.*, 2017), na previsão da doença de Parkinson (Vasquez-Correa *et al.*, 2016), no suporte de comunicação para pessoas com disfunção de fala (Balaji; Sadashivappa, 2015) e na avaliação da inteligibilidade de fala em pacientes com doenças

¹ <https://www.hapvida.com.br/>

bucais (Riemann *et al.*, 2016). No escopo do problema de documentação de prontuários e anamneses destaca-se os estudos de Liu *et al.* (2011), de Gür (2012), de Kodish-Wachs *et al.* (2018) e de Toledo *et al.* (2019), que buscaram verificar a viabilidade dessa tecnologia.

Esses estudos demonstram uma evolução significativa na precisão e na capacidade de transcrição do ASR, conforme evidenciado por métricas como *Word Error Rate* (WER), que mede a proporção de erros em uma transcrição em relação ao total de palavras faladas. Por exemplo, o estudo de Toledo *et al.* (2019) alcançou um WER de 12,30%, enquanto o estudo de Kodish-Wachs *et al.* (2018) obteve um WER de 35%. Apesar desses avanços, os sistemas de ASR frequentemente produzem transcrições brutas, sem processamento adicional, o que dificulta a identificação clara das informações apresentadas.

Uma abordagem para mitigar esses desafios é a integração de modelos ASR com Modelos Grandes de Linguagem, ou *Large Language Model* (LLM). Um LLM é um sistema avançado de inteligência artificial que processa e gera texto com fluência e coerência semelhantes às humanas (Naveed *et al.*, 2024). Existem diversos tipos de LLMs treinados para tarefas e contextos específicos, sendo os LLMs baseados em instrução, ou *Instruct-based*, um exemplo. Estes modelos são treinados para ajustar sua capacidade de generalização em novas tarefas (Yuan *et al.*, 2023), permitindo adaptação a diferentes domínios de texto e melhorando a eficácia em etapas de pós-processamento, como a transferência de estilo das transcrições geradas pelos ASR.

Além disso, os LLMs podem ser refinados para contextos específicos, como o médico, por meio de técnicas como *fine-tuning*. Este processo envolve o treinamento adicional de um modelo pré-treinado com dados específicos para melhorar seu desempenho em uma tarefa ou contexto (Naveed *et al.*, 2024). Exemplos de modelos de linguagem adaptados para o contexto biomédico são o BioGPT (Luo *et al.*, 2022) e o BioBERT (Lee *et al.*, 2019), que compartilham objetivos em comum de aprimorar a capacidade de processamento de linguagem natural no domínio biomédico. No entanto, esses modelos foram treinados utilizando, predominantemente, artigos científicos e textos biomédicos, o que restringe sua capacidade de compreender e processar informações específicas de anamneses médicas.

Assim, diante dos desafios enfrentados na documentação eficiente de anamneses médicas e as limitações encontradas em modelos de ASR, propõe-se o desenvolvimento e aplicação de um LLM como estratégia de pós-processamento para otimizar as transcrições de anamneses médicas. Essa abordagem busca não apenas reduzir o tempo dedicado à documentação da anamnese, mas também aprimorar a qualidade das informações registradas, contribuindo

para diagnósticos mais precisos e tratamentos mais eficazes. Os resultados deste estudo serão integrados ao projeto CEREIA, na linha 5 "Anamnese Assistida por IA", em que a aluna está participando.

O objetivo geral deste estudo é explorar o uso de LLMs como meio de aperfeiçoamento para transcrições de anamneses médicas. Especificamente, busca-se realizar fine-tuning de um LLM pré-treinado com dados de anamneses e avaliar a eficácia desses modelos na formatação e transferência de estilo das transcrições geradas pelos ASR. Espera-se que esse trabalho resulte em uma melhoria significativa na qualidade das informações registradas nas anamneses médicas, facilitando o processo de tomada de decisões clínicas e contribuindo para uma prática médica mais eficaz. A aplicação desses avanços no contexto do projeto CEREIA destaca a importância desses estudos para o aprimoramento das práticas médicas e da tecnologia assistiva na área da saúde.

Este trabalho também resultou na publicação de dois artigos: um artigo completo (Gonçalves *et al.*, 2024b) e um artigo demo (Gonçalves *et al.*, 2024a), apresentados no XXXIX Simpósio Brasileiro de Bancos de Dados (SBBDD).

2 REFERENCIAL TEÓRICO

Este capítulo explora os três temas principais para o embasamento teórico deste estudo. Na seção 2.1, são abordados o conceito de anamnese, sua relevância no contexto clínico e os elementos que compõem sua estrutura, além de particularidades do texto. A seção 2.2 discute o Reconhecimento Automático de Fala (ASR), destacando sua relação com a área de Processamento de Linguagem Natural (PLN), seus principais componentes, tendências atuais e desafios mais significativo. Por fim, a seção 2.3 aborda os principais conceitos relacionados a Modelos Grandes de Linguagem, apresentando os principais conceitos, estratégias e as tendências atuais.

2.1 Anamnese Médica

A Semiologia é uma área da medicina que estuda os sinais e sintomas das doenças, visando à identificação e compreensão das condições de saúde dos pacientes. Dentro desse contexto, a anamnese surge como "o instrumento de coleta da histórica clínica e dos temas relativos à queixa principal do paciente"(Yehia *et al.*, 2024). Geralmente, ela é construída durante o Exame Clínico do paciente, uma abordagem fundamentada na coleta sistemática dos dados (Yoshikawa; Castro, 2015, p. 15).

Porém, a anamnese é muito mais que apenas um documento de coleta. Segundo Yehia *et al.* (2024), ela é "responsável pela criação do vínculo entre o médico e o paciente", sendo fundamental à prática médica. Ela deve ser construída em conjunto ao paciente, onde o médico deve orientá-lo e construir um ambiente no qual ele se abra em relação a sua própria doença, o que se sente, de que forma e há quanto tempo, além de outras informações individuais fora a doença em si (Soares *et al.*, 2016).

Essa prática não apenas orienta o processo diagnóstico, mas também contribui para a personalização do atendimento, reconhecendo o paciente como um indivíduo com história, valores e experiências (Yehia *et al.*, 2024). Ao explorar aspectos sociais, psicológicos e biográficos do paciente, assim como suas ideias, preocupações e expectativas, o profissional de saúde pode identificar fatores que influenciam diretamente na sua condição clínica, aproximando mais o médico de um raciocínio clínico bem embasado e uma hipótese diagnóstica (Benseñor, 2013; Yoshikawa; Castro, 2015, p. 15).

Ademais, a qualidade da anamnese reflete diretamente na qualidade da relação

médico-paciente estabelecida. Quando construída de forma clara e organizada, ela se torna a principal fonte de informações para a construção do diagnóstico provável o que leva a uma melhor gestão de recursos, seja do hospital como do próprio paciente, sendo responsável por, aproximadamente, 80% dos diagnósticos (Benseñor, 2013). Assim, o respeito com paciente, empatia, escuta focada e outras habilidades sociais se tornam algo que deve fazer parte da formação e dia a dia do médico, fortalecendo o vínculo e auxiliando no tratamento do paciente (Fuehrer *et al.*, 2024).

2.1.1 Estrutura Base das Anamneses

As anamneses podem ser elaboradas por diversos motivos além de documentar o diagnóstico e tratamento do paciente. Elas podem servir como auxílio à memória do médico, ser utilizadas por outros profissionais de saúde envolvidos no cuidado ao paciente e também registrar todo o processo de assistência médica por razões legais (Dalianis, 2018, p. 8). Por isso, esse documento precisa ser claro e bem organizado, promovendo "uma boa comunicação entre todos aqueles que participam da assistência ao paciente, ajudando a coordenar suas atividades"(Yoshikawa; Castro, 2015).

Inicialmente, as anamneses eram registradas em papel, permitindo a percepção visual de informações como a primeira visita do paciente ou a mais recente, já que o papel ficava amarelado com o tempo (Dalianis, 2018, p. 10). Com a digitalização, esses documentos passaram a ser registrados eletronicamente. De acordo com Garrett *et al.* (1986) e Mata *et al.* (2015), o registro eletrônico de anamneses reduziu significativamente o tempo de documentação, melhorou a disponibilidade de dados e diminuiu os custos. Independentemente do meio no qual são registradas, as anamneses seguem uma **estrutura base**.

Essa estrutura é essencial tanto para o registro quanto para guiar a construção da história clínica, orientando as perguntas durante o exame e a entrevista clínica (Machado *et al.*, 2018). Uma anamnese clássica é composta pelos seguintes tópicos: Identificação, Queixa Principal (QP), História da Doença Atual (HDA), História Patológica Progressiva (HPP), História Pessoal e Social (HPS), História Familiar (HF), Revisão de Sistemas (RS), Exame(s) Físico(s), Hipótese Diagnóstica e Conduta (Machado *et al.*, 2018; Yoshikawa; Castro, 2015).

Os dois primeiros tópicos representam a fase inicial da entrevista e o estabelecimento do contato entre médico e paciente. A **Identificação** reúne os dados pessoais do paciente, como nome, idade, sexo, estado civil, nacionalidade, profissão, residência, religião, entre outros.

Yoshikawa e Castro (2015) recomendam que a identificação seja feita de forma completa. A **Queixa Principal (QP)**, também chamada de "Motivo da Consulta" ou "Agenda do Paciente", registra o principal motivo da busca por atendimento. Recomenda-se que seja escrita com as palavras do paciente (Yoshikawa; Castro, 2015, p. 17), representando a "razão real para o paciente ter buscado auxílio" (Machado *et al.*, 2018). Esse tópico serve como ponto de partida para a conversa, permitindo que o paciente exponha sua situação.

A **História da Doença Atual (HDA)**, ou História da Moléstia Atual, consiste na exploração detalhada do tópico anterior. Essa fase é caracterizada por uma narrativa espontânea do paciente, em que devem ser observadas pistas emocionais e o impacto da doença na sua vida (Machado *et al.*, 2018). Na documentação, a HDA apresenta-se como uma "narrativa clara, sistemática e sintética, destacando as relações cronológicas e os sintomas atuais do paciente" (Yoshikawa; Castro, 2015, p. 17).

Após a HDA, seguem os tópicos que abordam a história do paciente: **História Patológica Progressiva (HPP)**, **História Pessoal e Social (HPS)** e **História Familiar (HF)**. Esses aspectos podem se sobrepor e, às vezes, é difícil separá-los (Machado *et al.*, 2018). A HPP explora comorbidades, histórico gestacional, condições de nascimento e desenvolvimento, vacinas e alergias. A HPS examina a rotina do paciente, incluindo dieta, hábitos (exercícios, consumo de álcool, higiene, sono), relações afetivas e de trabalho, *hobbies* e condições de moradia. Já a HF investiga doenças hereditárias, como diabetes e câncer, além da incidência de condições comuns na família (Machado *et al.*, 2018; Yoshikawa; Castro, 2015, p. 18). Dessa forma, enquanto a HPP e a HF focam na doença, a HPS concentra-se no indivíduo e no seu contexto.

A Revisão de Sistemas (RS) e os Exames Físicos representam a fase de conclusão, direcionando a investigação para a análise dos problemas. A **RS** revisa sintomas relacionados à queixa do paciente que podem ter sido omitidos ou esquecidos nas etapas anteriores (Yoshikawa; Castro, 2015, p. 18). As perguntas começam de forma ampla e, em seguida, tornam-se mais específicas (Machado *et al.*, 2018). Já os **Exames Físicos** auxiliam na confirmação das hipóteses, identificando sinais não percebidos pelo paciente. Esses exames exigem habilidade técnica do médico e devem ser realizados sistematicamente, abrangendo áreas relevantes para a queixa (Yoshikawa; Castro, 2015, p. 21). Juntos, a RS e os exames físicos conectam os achados clínicos às informações coletadas na entrevista.

Por fim, a anamnese é concluída com a Hipótese Diagnóstica e a Conduta. A

Hipótese Diagnóstica reflete o raciocínio clínico do médico, baseado nas informações subjetivas da entrevista e nos dados objetivos dos exames. Ela deve ser clara, justificada e considerar as possibilidades mais prováveis (Yoshikawa; Castro, 2015, p. 22). A **Conduta**, por sua vez, descreve as ações propostas, como solicitação de exames complementares, prescrição de medicamentos ou encaminhamentos (Yoshikawa; Castro, 2015, p. 22).

Essa estrutura garante que a anamnese seja abrangente e funcional, tanto para o diagnóstico quanto para a documentação e continuidade do cuidado ao paciente. Contudo, ela pode variar conforme a área de aplicação, o médico ou a instituição, adaptando-se a situações específicas.

2.1.2 Particularidades do Texto e Desafios

Os textos das anamneses médicas possuem características únicas que dificultam sua análise, interpretação e padronização. Isso ocorre porque são frequentemente apresentados em linguagem médica especializada, repleta de termos técnicos, jargões e abreviaturas que podem variar entre diferentes disciplinas ou profissionais de saúde (Caseli; Nunes, 2024, p. 592). Essa variação pode resultar em incompreensões, até mesmo entre profissionais da área médica (Dalianis, 2018, p. 23).

Além disso, dependendo do contexto de coleta, os registros médicos podem conter dados estruturados, semiestruturados e não estruturados, sendo que as narrativas em texto livre (dados não estruturados) são as mais desafiadoras para análise (Caseli; Nunes, 2024, p. 592). Essas narrativas frequentemente apresentam características como:

- **Erros linguísticos e gramaticais:** até 10% dos textos em prontuários médicos contêm erros (Dalianis, 2018, p. 25).
- **Uso de abreviaturas e acrônimos:** cerca de 10% do texto pode incluir abreviaturas, e até 5% acrônimos (Dalianis, 2018, p. 34).
- **Ausência de frases completas:** muitas vezes, as narrativas são compostas por fragmentos ou expressões telegráficas (Dalianis, 2018, p. 21).

Embora esses fatores possam dificultar o uso automatizado dessas informações, tecnologias de Processamento de Linguagem Natural (PLN) têm sido empregadas para auxiliar na análise dos textos médicos. No entanto, destaca-se que essas tecnologias devem complementar e melhorar o cuidado ao paciente, e não substituir a expertise dos profissionais de saúde (Caseli; Nunes, 2024, p. 598).

Essas particularidades do texto médico refletem não apenas os desafios inerentes ao registro e à comunicação entre profissionais, mas também a necessidade de maior padronização e de ferramentas tecnológicas que contribuam para a qualidade do atendimento.

2.2 *Automatic Speech Recognition*

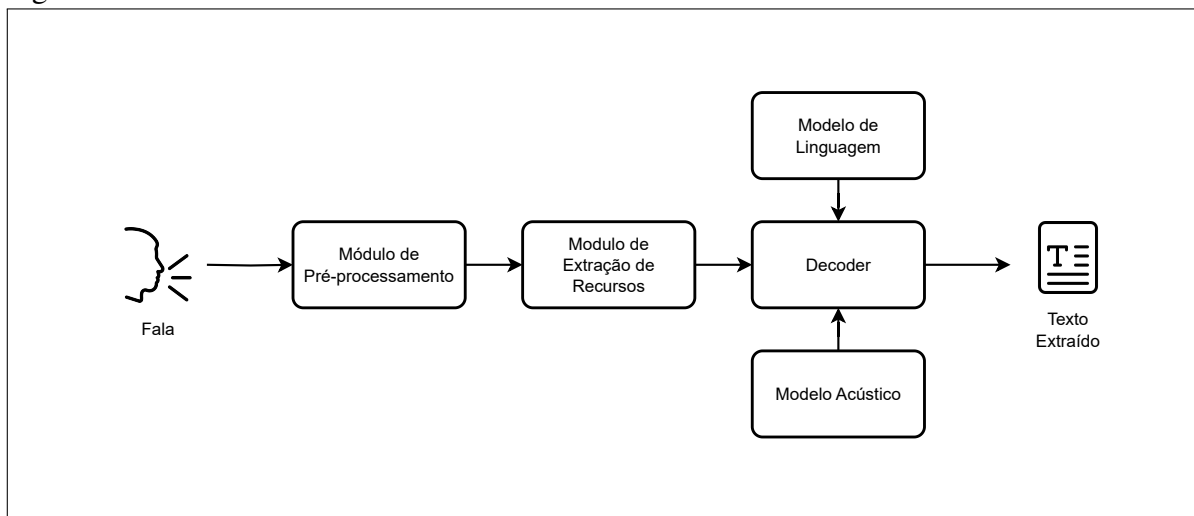
A fala é algo inerente à existência humana. Segundo Sapir (1921, p.10), "não há fato mais surpreendente sobre a linguagem do que a sua universalidade". Diante disso, Malik *et al.* (2021) afirma que pode-se presumir que as pessoas se sentem mais confortáveis usando a fala como meio de interação com computadores e máquinas do que com outros meios. Nesse contexto, surgem os sistemas de reconhecimento de fala, ou ASRs.

De forma geral, um ASR é um sistema que recebe uma onda sonora e a transforma em texto (Malik *et al.*, 2021). Embora seu principal objetivo seja a transcrição da fala, um ASR também pode ser utilizado para identificar a língua que está sendo falada, detectar emoções e reconhecer a identidade do locutor (O'Shaughnessy, 2008). Essas funcionalidades são essenciais para a área de PLN, que busca desenvolver métodos para a compreensão e manipulação da linguagem humana por meio de sistemas computacionais (Caseli; Nunes, 2024, p. 10).

A estrutura básica de um ASR, como ilustrado na Figura 1, pode ser dividida em cinco módulos: módulo de pré-processamento, módulo de extração, modelo de linguagem, modelo acústico e módulo de *decoder*. O objetivo do **módulo de pré-processamento** é limpar o áudio que é recebido como entrada. Como nem sempre essa entrada é gravada em um ambiente silencioso ou sem interferências externas, é necessário reduzir ao máximo o barulho que pode estar no sinal, ou seja, reduzir a relação sinal-ruído. Para isso, podem ser aplicado vários filtros ou métodos diferentes. Em seguida, o sinal filtrado é enviado para o **módulo de extração**. Essa fase geralmente é feita em três etapas: análise do sinal, onde é feita uma análise espectral temporal para gerar características brutas; formação de um vetor de características expandido, que inclui características estáticas e dinâmicas; e, por fim, a transformação desses vetores em formas mais compactas e robustas (Karpagavalli; Chandra, 2016).

O módulo do **modelo de linguagem** é responsável por fornecer o contexto necessário para a interpretação dos fonemas previstos pelo ASR. Ele é composto por regras e estatísticas que definem quais sequências de palavras são válidas e como elas podem ocorrer (Karpagavalli; Chandra, 2016). Esse modelo melhora a precisão do reconhecimento ao prever palavras e frases com base nos fonemas identificados. Embora muitos ASRs possam funcionar sem um

Figura 1 – Estrutura básica de um ASR



Fonte: Elaborada pela autora, baseado em Karpagavalli e Chandra (2016) e Malik *et al.* (2021).

modelo de linguagem, a sua inclusão geralmente aumenta a eficiência do sistema (Malik *et al.*, 2021). Já o módulo do **modelo acústico** representa as características acústicas dos fonemas a serem reconhecidos (Karpagavalli; Chandra, 2016). Esse modelo é construído a partir de uma grande base de dados de fala, utilizando algoritmos de treinamento para criar representações estatísticas para cada fonema em uma língua. O modelo acústico descreve a probabilidade de uma observação específica dada um fonema, ajudando o decodificador de fala a identificar os sons distintos produzidos pelo usuário (Karpagavalli; Chandra, 2016). Por fim, o **módulo de decoder** determina a sequência de palavras mais provável, ou o *decoding*, com base nas observações e nos modelos acústico e de linguagem.

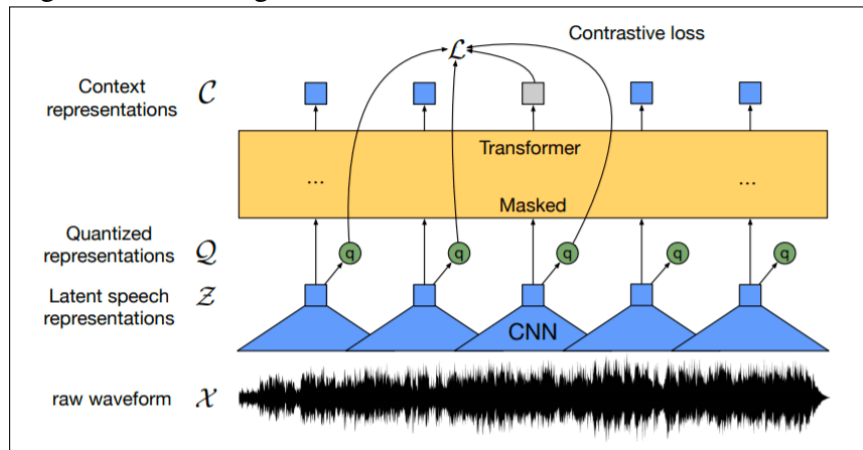
2.2.1 Modelos do Estado da Arte

O **Wav2Vec 2.0** (Baevski *et al.*, 2020) foi desenvolvido pela MetaAI¹ para lidar com o desafio de treinar modelos ASR, especialmente quando há poucos dados rotulados disponíveis. Em vez de exigir extensos conjuntos de dados rotulados, ele aprende unidades de fala discretas juntamente com representações contextualizadas. A arquitetura do modelo inclui um codificador de características que processa o áudio bruto com convoluções temporais, normalização de camadas e uma função de ativação GELU. A saída desse codificador é então enviada para uma rede de contexto baseada na arquitetura Transformer (Vaswani *et al.*, 2017) (Figura 2).

Essa abordagem permite que o Wav2Vec 2.0 obtenha representações vetoriais diretamente do áudio sem a necessidade de extração de atributos acústicos, abordando a escassez de

¹ <https://ai.meta.com/>

Figura 2 – Abordagem do Wav2Vec 2.0



Fonte: Baevski *et al.* (2020)

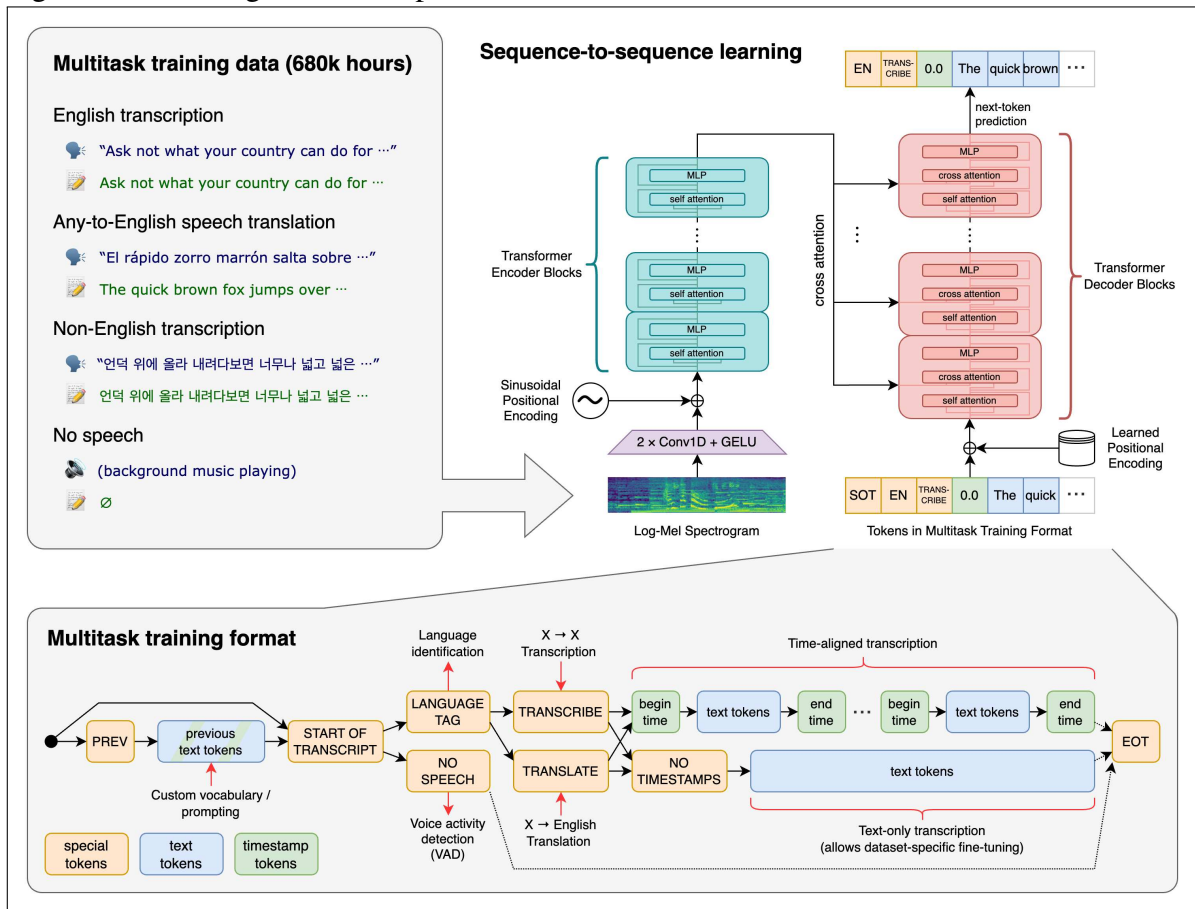
dados rotulados, especialmente em línguas com poucos recursos. O modelo se destaca por sua eficiência, precisando de muito menos áudio de treinamento em comparação com arquiteturas tradicionais, o que é uma vantagem significativa mesmo para línguas bem representadas como o inglês.

De maneira complementar, o **Whisper** (Radford *et al.*, 2023), desenvolvido pela OpenAI², foi projetado para superar as limitações dos codificadores de áudio pré-treinados não supervisionados, como o Wav2Vec 2.0. Enquanto esses codificadores são eficazes em aprender representações de fala a partir de áudio bruto, eles carecem de um decodificador de desempenho comparável, necessitando de ajuste fino para tarefas específicas como o reconhecimento automático de fala (ASR). Whisper resolve essa limitação ao introduzir uma arquitetura Transformer de codificador-decodificador, ou *encoder-decoder*, que utiliza modelos de sequência para sequência (seq2seq) para prever transcrições sem necessidade de extensa padronização (Figura 3).

Além disso, o Whisper realiza tarefas como identificação de idioma e tradução para o inglês. Treinado com 680.000 horas de dados em 96 idiomas, Whisper se adapta bem a outros conjuntos de dados sem precisar de *fine-tuning* específico para cada conjunto, tornando-o um modelo versátil e eficiente para múltiplas tarefas e idiomas. Dessa forma, tanto o Wav2Vec 2.0 quanto o Whisper apresentam avanços significativos na área de reconhecimento de fala, cada um abordando diferentes aspectos e desafios deste campo.

² <https://openai.com/>

Figura 3 – Abordagem do Whisper



Fonte: Radford *et al.* (2023)

2.2.2 Desafios

Mesmo com os avanços, ainda existem desafios e limitações na criação e uso de ASR. A fala é uma atividade humana que varia de um grupo social para outro (Sapir, 1921). Segundo Karpagavalli e Chandra (2016), as variações mais notáveis são no estilo de fala, diferentes dialetos e sotaques, além das variações que podem surgir quando levamos gênero, cadência e velocidade em consideração. Todas essas possíveis características e variações são alguns dos maiores obstáculos no desenvolvimento de ASRs através dos anos, como afirmam O'Shaughnessy (2008) e Malik *et al.* (2021).

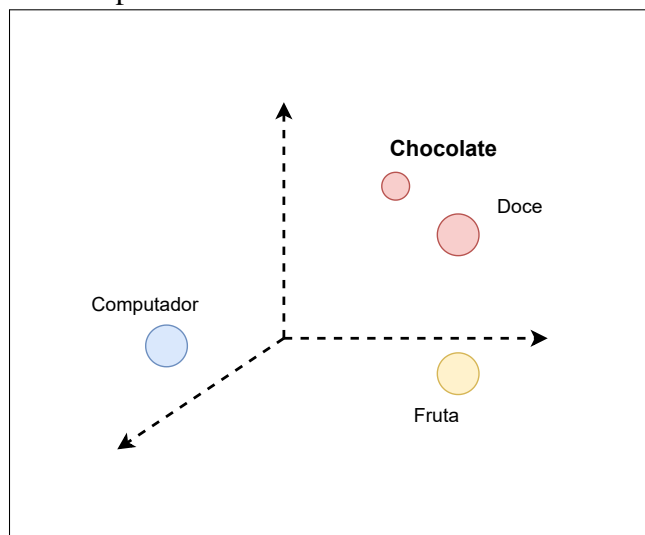
Além disso, muitas vezes a saída desses modelos é um texto corrido, sem formatação ou estilização. Isso pode dificultar a compreensão e a utilização das transcrições em contextos específicos, onde a organização e estrutura do texto são essenciais, como uma anamnese. Consequentemente, aprimorar a saída dos sistemas de ASR com técnicas de formatação torna-se uma etapa crucial para aumentar a eficácia e a usabilidade das transcrições geradas. Nesse sentido, a aplicação de LLMs como ferramenta de pós-processamento pode não apenas corrigir erros na

transcrição, mas também estruturar o texto de maneira mais organizada e estilizada, atendendo melhor às necessidades dos usuários finais.

2.3 Modelos Grandes de Linguagem

Assim como ASRs transformam sinais sonoros em texto para permitir que máquinas compreendam a fala humana, a compreensão textual por máquinas também depende da conversão de texto em representações numéricas. Como já mencionado na seção 2.2, a linguagem é essencial para a comunicação e expressão humana, assim, além da simples representação numérica, é necessário que esse formato numérico denote alguma conotação semântica e não apenas uma identificação (Caseli; Nunes, 2024, p. 317). Para alcançar isso, modelos de linguagem utilizam representações vetoriais, que podem ter uma ou mais dimensões, e esse processo pode ser realizado através de diferentes algoritmos. Esse processo é conhecido como *Word Embedding* e pode ser feito através de diferentes algoritmos, como o Word2Vec. Com essas representações, são realizados cálculos para determinar a similaridade semântica entre palavras e conceitos (Figura 4) e os modelos de linguagem utilizam essas representações para tarefas como tradução, resumo, recuperação de informações, interações conversacionais (Naveed *et al.*, 2024).

Figura 4 – Similaridade Semântica entre 'chocolate' e outras palavras



Fonte: Elaborada pela autora.

De acordo com Caseli e Nunes (2024, p. 318), um modelo de linguagem, ou *Language Model* (LM), pode ser definido como uma representação computacional simplificada da língua. Seu objetivo principal é identificar a distribuição de probabilidade das palavras em um

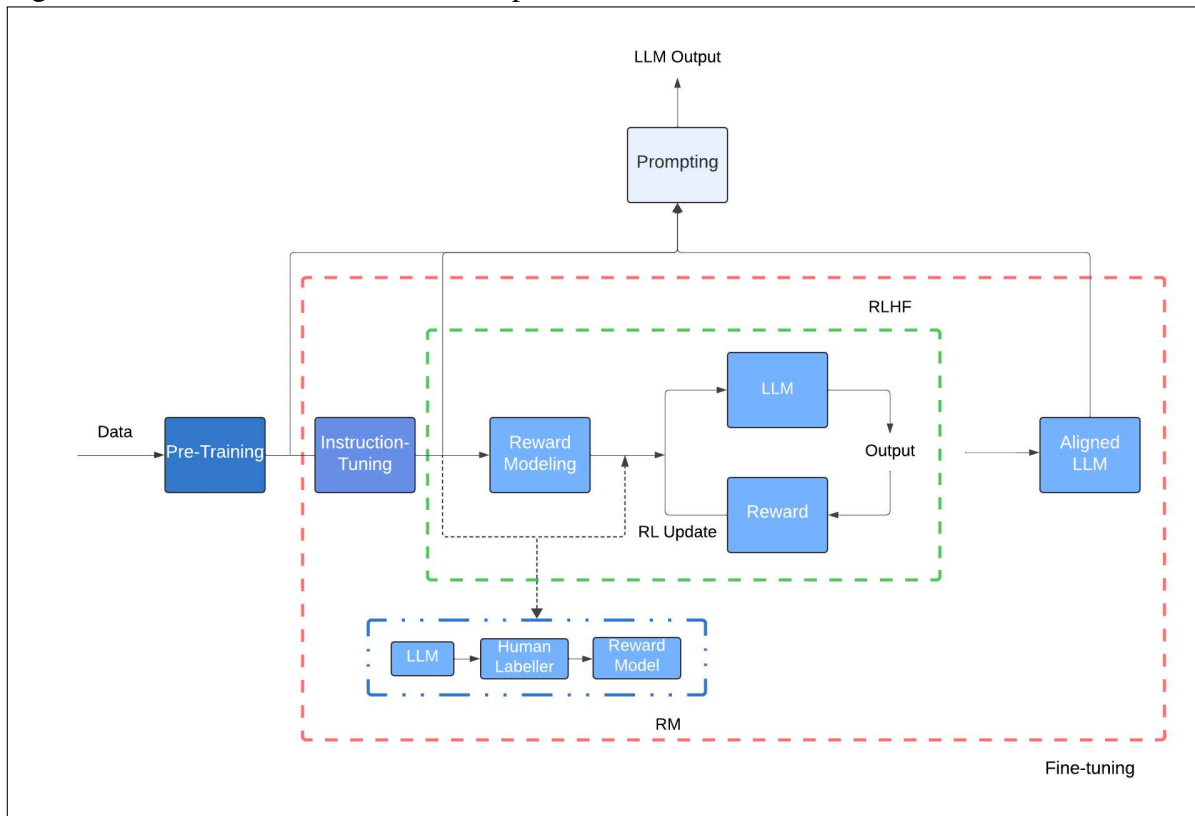
idioma, o que permite ao modelo gerar textos novos, completar sentenças e prever a probabilidade de diferentes sequências de palavras. A evolução desses modelos começou com técnicas estatísticas simples, como os modelos n-gram, e progrediu para métodos mais avançados, como os modelos baseados em *deep learning* (Hadi *et al.*, 2023; Caseli; Nunes, 2024). Recentemente, avanços significativos foram observados nos modelos de linguagem, atribuídos principalmente a arquitetura Transformer, proposta por Vaswani *et al.* (2017), ao aumento das capacidades computacionais e à disponibilidade de grandes conjuntos de dados de treinamento (Naveed *et al.*, 2024).

Nesse contexto, surgem os LLMs. Um LLM é um modelo de linguagem neural que se diferencia dos demais modelos de linguagem pela sua quantidade enorme de parâmetros, seu enquadramento como métodos de IA generativa e suas habilidades emergentes, como o aprendizado em contexto e a estratégia de cadeia de pensamento, ou *Chain-of-Thought* (CoT) (Caseli; Nunes, 2024, p. 357-358). Essas características permitem que os LLMs realizem tarefas com níveis de desempenho comparáveis aos humanos (Naveed *et al.*, 2024). Uma visão geral do fluxo de etapas de um LLM pode ser vista na Figura 5, partindo de um modelo pré-treinado, que pode ser ajustado ainda mais para uma tarefa específica ou para um contexto específico, ou utilizado diretamente.

Os LLMs podem ser baseados em diversas arquiteturas Transformer, cada uma oferecendo características distintas para o processamento de linguagem. A arquitetura ***Encoder-Decoder*** processa as entradas através do codificador, que utiliza autoatenção para compreender a sequência completa, e passa a representação intermediária para o decodificador, que gera a saída utilizando atenção cruzada e processa a sequência passo a passo. Em contraste, a arquitetura ***Causal Decoder*** não possui um codificador e gera a saída com base exclusivamente nos tokens anteriores, garantindo que a previsão de cada token dependa apenas dos tokens que vieram antes dele. A arquitetura ***Prefix Decoder*** permite uma atenção bidirecional, onde a atenção não está restrita às informações passadas, proporcionando uma análise mais abrangente da sequência.

Finalmente, o modelo ***Mixture-of-Experts*** (MoE) é uma variante que incorpora especialistas paralelos e independentes, com um roteador responsável por direcionar tokens para esses especialistas. Esta abordagem é eficiente e permite o aumento do tamanho do modelo sem um incremento proporcional no custo computacional, ativando apenas alguns especialistas por vez, o que resulta em um desempenho comparável aos modelos densos (Naveed *et al.*, 2024).

Figura 5 – Visão Geral do Fluxo de Etapas de um LLM



Fonte: Naveed *et al.* (2024)

2.3.1 Prompt Engineering

Os LLMs utilizam *prompts* para guiar a geração de texto e realizar tarefas específicas. Um *prompt* pode ser definido como "um texto em linguagem natural (em oposição à linguagem de programação) que especifica uma instrução do que deve ser feito" (Caseli; Nunes, 2024, p. 345). Embora não haja uma definição formal, *prompt engineering* refere-se ao design e à redação de *prompts* de forma a obter uma resposta desejada dos LLMs (Hadi *et al.*, 2023). A formulação dos *prompts* é crucial, pois pode influenciar significativamente a qualidade e a relevância das respostas geradas pelo modelo. No fluxo de etapas de um LLM, como ilustrado na Figura 5, o módulo de *prompting* é responsável por receber essas instruções e então o modelo gera a saída.

Existem algumas técnicas que podem ser utilizadas para extrair a melhor saída dado um determinado objetivo. Os LLMs podem ser adaptados às instruções sem a necessidade de *fine-tuning*³ ou, em alguns casos, com *fine-tuning* em dados que contêm diferentes estilos de *prompts* (Naveed *et al.*, 2024). No **zero-shot prompting**, os LLMs são capazes de responder a

³ *Fine-tuning*, ou ajuste fino, é o processo de ajustar um modelo pré-treinado em um conjunto de dados específico para melhorar seu desempenho em uma tarefa particular (Church *et al.*, 2021).

consultas nunca vistas antes, sem a necessidade de exemplos no *prompt*. ***In-context learning***, também conhecido como ***few-shot learning***, múltiplos pares de entrada-saída são apresentados ao modelo para gerar a resposta desejada. Os LLMs também podem ser induzidos a gerar respostas para problemas lógicos e planejamento de tarefas com diferentes estilos de *prompting*, conhecido como ***Reasoning in LLMs***. No ***Single-Turn Instructions***, os LLMs são consultados uma única vez com todas as informações relevantes no *prompt*. Por fim, no ***Multi-Turn Instructions***, resolver uma tarefa complexa requer múltiplas interações com os LLMs, onde o feedback e as respostas são usados como entrada para rodadas subsequentes (Naveed *et al.*, 2024; Hadi *et al.*, 2023).

Adicionalmente, técnicas como o ***Negative Prompting*** direcionam o LLM sobre aspectos que devem ser evitados ou excluídos durante o processo de geração, enquanto o ***Visual Prompting*** utiliza *prompts* visuais, como imagens ou música, além de texto, para fornecer ao modelo um ponto de partida ou referência para uma tarefa gerativa (Hadi *et al.*, 2023). Essas técnicas ajudam a personalizar e a refinar as respostas geradas pelos LLMs, melhorando sua eficácia em diversas aplicações.

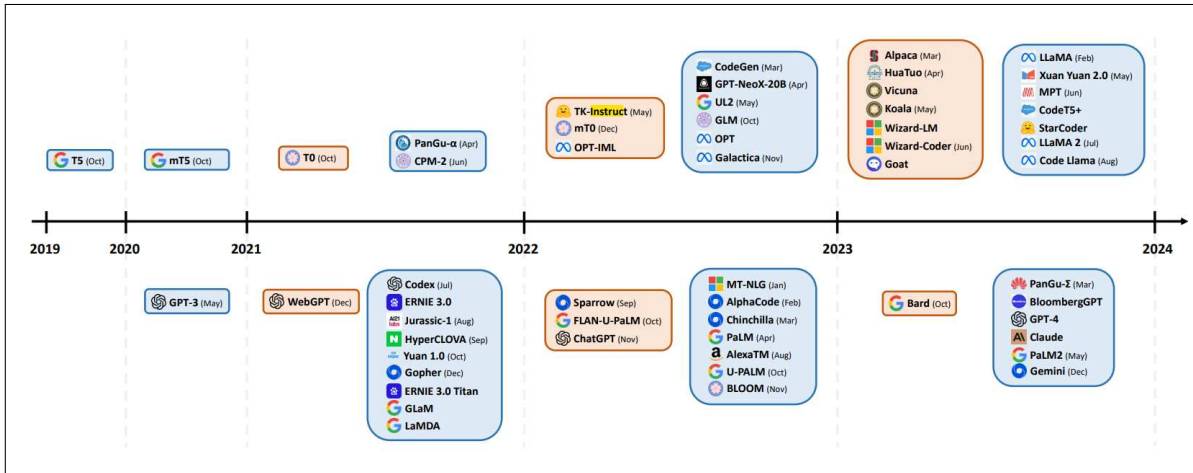
2.3.2 ***Instruct Based Models***

Modelos baseados em instrução surgem a partir da necessidade de se responder de forma eficaz às consultas do usuário, possibilitando que o modelo responda adequadamente às perguntas do usuário (Naveed *et al.*, 2024). Para que isso aconteça, é realizado o processo de *instruction-tuning*, que é um *fine-tuning* baseado em instruções. Nesse processo, o modelo pré-treinado é ajustado usando dados formatados com instruções, incluindo uma instrução e um par de entrada-saída. O *instruction-tuning* melhora a generalização zero-shot e o desempenho em tarefas subsequentes (Naveed *et al.*, 2024). A importância dessa abordagem é evidenciada pela evolução dos modelos ao longo do tempo, como ilustrado na Figura 6.

2.3.3 ***Transferência de Estilo Textual***

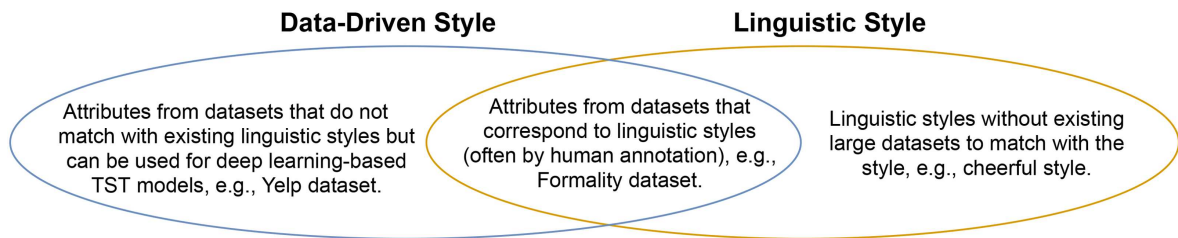
A Transferência de Estilo Textual (TST), ou *Text Style Transfer*, é uma tarefa em PLN que envolve a modificação de um texto para alterar seu estilo enquanto mantém seu conteúdo semântico (Mukherjee; Dušek, 2024). Essa tarefa tem uma ampla gama de aplicações, como a adaptação de tons de escrita, a simplificação de textos, a transformação de formal para informal e a geração de conteúdo estilizado (Jin *et al.*, 2022; Mukherjee; Dušek, 2024).

Figura 6 – LLMs lançados a partir de 2019 em ordem cronológica: modelos em azul representam ‘pré-treinados’, enquanto modelos em laranja são ‘instruction-tuned’



Fonte: Naveed *et al.* (2024)

Figura 7 – Diagrama de Venn da definição linguística de estilo e definição de estilo baseada em dados.



Fonte: Jin *et al.* (2022)

No contexto da TST, o "estilo" de um texto pode ser definido a partir de duas abordagens principais: a definição linguística e a definição orientada por dados (Figura 7). Na abordagem linguística, estilo refere-se às características inerentes a cada enunciado, incluindo o uso de dispositivos estilísticos como metáforas, escolha lexical e estruturas sintáticas, além de aspectos de nível discursivo (McDonald; Pustejovsky, 1985). Por outro lado, a definição orientada por dados considera o estilo como atributos gerais no texto que variam entre conjuntos de dados, como formalidade, tom ou polaridade de sentimento. Essa abordagem é amplamente utilizada em métodos baseados em aprendizado profundo, que dependem de grandes corpora para aprender estilos (Jin *et al.*, 2022).

LLMs têm desempenhado um papel central em pesquisas recentes sobre TST. Entre as técnicas baseadas em LLMs estão o uso de *prompts* em configurações *zero-shot* e *few-shot*, além do *fine-tuning* em modelos base. Embora os *prompts* ofereçam soluções rápidas e flexíveis, o *fine-tuning* frequentemente resulta em melhorias significativas na performance, especialmente em tarefas específicas, como transferência de sentimentos (Mukherjee *et al.*, 2024). Estudos recentes também exploram o uso de neurônios específicos de estilo para guiar os LLMs na realização da

TST, desativando neurônios relacionados ao estilo original para aumentar a precisão do texto gerado (Lai *et al.*, 2024).

Apesar dos avanços, a TST apresenta desafios consideráveis. Primeiro, não há separação clara entre estilo e conteúdo, dificultando o controle preciso sobre os atributos estilísticos do texto (Mukherjee; Dušek, 2024). Segundo, faltam dados paralelos⁴ para treinar modelos de forma supervisionada, especialmente em domínios especializados. Terceiro, a avaliação de resultados em TST carece de métricas padronizadas, sendo frequentemente realizada por meio de medidas automáticas, como BLEU (Papineni *et al.*, 2002), e avaliações humanas, que verificam fluência, consistência semântica e intensidade da transformação estilística (Jin *et al.*, 2022).

A TST encontra aplicações práticas em diversas áreas, como geração de diálogos com personas consistentes, personalização de traduções automáticas para diferentes estilos, anonimização de identidade textual e simplificação de texto (Mukherjee; Dušek, 2024; Jin *et al.*, 2022). A adaptação de modelos de linguagem para diferentes domínios é um aspecto crítico em tarefas de TST, pois as características do estilo podem variar drasticamente de acordo com o público-alvo ou o contexto de aplicação.

Em ambientes especializados, como o domínio médico, a TST enfrenta obstáculos adicionais. A documentação médica, por exemplo, exige não apenas a preservação do conteúdo semântico, mas também a precisão e a clareza necessárias para garantir que os profissionais de saúde possam fazer uso adequado da informação. Além das particularidades e variações já mencionadas na subseção 2.1.2.

2.3.4 Modelos do Estado da Arte

Como mencionado, a arquitetura *Transformer* revolucionou a área de LMs. A seguir, são apresentados três modelos que representam o estado da arte atual dos LLMs.

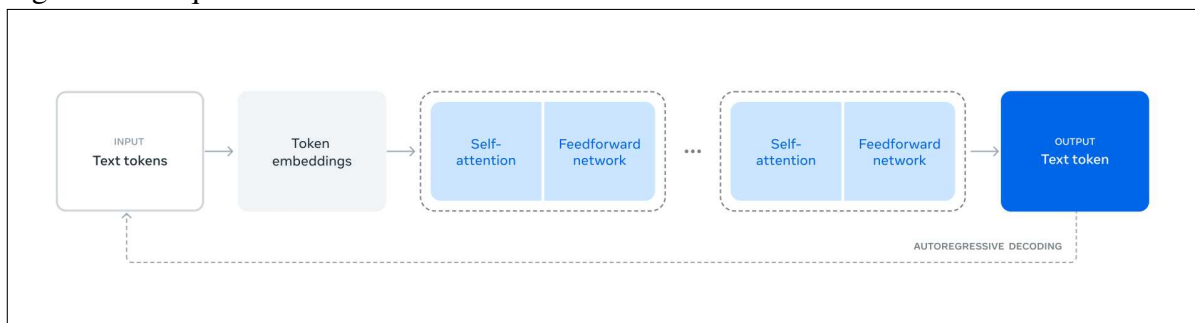
O primeiro modelo é o **Mistral 7B** (Jiang *et al.*, 2023), desenvolvido pelo Mistral AI. Com 7 bilhões de parâmetros, este modelo é projetado para oferecer alto desempenho e eficiência. Baseado na arquitetura *Transformer*, o Mistral 7B utiliza mecanismos avançados de atenção, como o *Grouped-Query Attention* (GQA) e o *Sliding Window Attention* (SWA), para otimizar a velocidade de inferência e reduzir a demanda de memória durante a decodificação.

⁴ Dados paralelos referem-se a conjuntos de dados que consistem em pares de entradas correspondentes em diferentes estilos, permitindo que o modelo aprenda as transformações estilísticas diretamente. Um exemplo seria um par de frases, uma escrita de forma formal e a outra informal, ambas com o mesmo conteúdo semântico (Mukherjee; Dušek, 2024).

Isso permite tamanhos de lote maiores e maior *throughput*, essenciais para aplicações em tempo real. O Mistral 7B também possui uma variação 8B, com 8 bilhões de parâmetros, e versões baseadas em instrução.

O **Llama 3**⁵ é o modelo de linguagem mais recente desenvolvido pela MetaAI, baseado na arquitetura *Transformer*. Com 405 bilhões de parâmetros, o Llama 3 suporta uma janela de contexto de até 128.000 tokens. O treinamento e a arquitetura simplificados do Llama 3 podem ser vistos na Figura 8.

Figura 8 – Arquitetura e treinamento do Llama 3



Fonte: MetaAI (2024)

Durante o pré-treinamento, o modelo é exposto a um corpus multilíngue de 15,6 trilhões de *tokens*, inicialmente com uma janela de contexto de 8.000 *tokens*, que depois é expandida para 128.000 *tokens*. No pós-treinamento, o modelo é ajustado com *feedback* humano por meio de técnicas de *fine-tuning* supervisionado e otimização direta de preferências. Novas capacidades, como o uso de ferramentas e medidas de segurança, são integradas. Além disso, o Llama 3 inclui módulos multimodais, com codificadores de imagem e fala treinados separadamente e integrados ao modelo por adaptadores para alinhar informações visuais e de fala com a linguagem.

Por fim, o **Phi3** (Abdin *et al.*, 2024) é um modelo desenvolvido pela Microsoft, focado em eficiência e execução em dispositivos móveis. Também baseado na arquitetura *Transformer*, o phi-3 inclui variantes como phi-3-mini (3,8 bilhões de parâmetros), phi-3-small (7 bilhões de parâmetros) e phi-3-medium (14 bilhões de parâmetros). Durante o treinamento, o modelo é exposto a 3,3 trilhões de tokens, combinando dados públicos filtrados e sintéticos.

O treinamento é realizado em duas fases principais: *Supervised Fine-Tuning*, que utiliza um conjunto diversificado de dados para várias tarefas e domínios, e *Direct Preference Optimization*, que ajusta o modelo para melhorar seu comportamento em interações de chat.

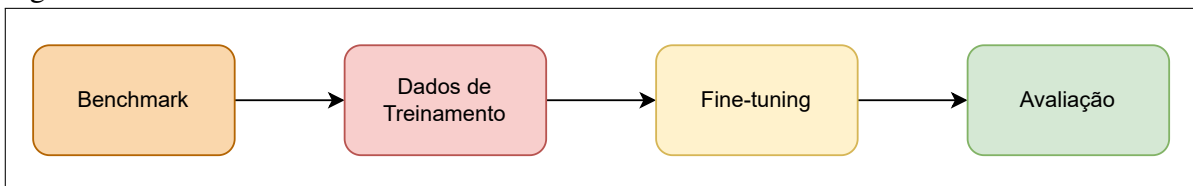
⁵ <https://ai.meta.com/research/publications/the-llama-3-herd-of-models/>

A variante phi-3-vision é projetada para processar entradas multimodais, combinando texto e imagens, utilizando um codificador de imagem baseado na arquitetura CLIP ViT-L/14 e um decodificador de *Transformer*, permitindo que o modelo realize tarefas de raciocínio visual e textual.

3 METODOLOGIA

O presente trabalho propõe o desenvolvimento de um LLM especializado na área médica, com foco em aprimorar anamneses médicas. A metodologia adotada para alcançar tal objetivo pode ser dividida em quatro etapas: *Benchmark*, Coleta e Geração de dados de Treinamento, *Fine-tuning* e Avaliação (Figura 9).

Figura 9 – Processo de Desenvolvimento



Fonte: Elaborada pela autora.

3.1 Benchmark

Visando avaliar a eficiência de diferentes LLMs na tarefa de formatação de anamneses médicas, foi realizado um *benchmark*. Essa etapa permite comparar diretamente o desempenho de cada modelo, identificando aquele capaz de produzir as formatações mais claras, consistentes e que evitem alucinação, ou seja, inserção de conteúdos não originalmente presentes, dada a necessidade de confiança e a natureza sensível do escopo do problema.

3.1.1 Elaboração do Prompt

Primeiramente, elaborou-se um *prompt* padrão que será utilizado para geração de todas as formatações. Para isso, foram utilizadas duas técnicas de prompt engineering: *Single-Turn Instruction*, que fornece todas as informações necessárias para a tarefa, e *Negative Prompt*, as quais limitam o que o modelo *não* deve fazer. Ademais, o LLM foi instruído a se comportar como um "formatador de prontuários e anamneses médicas em português".

Neste prompt, incluiu-se uma breve descrição do que é uma anamnese, bem como dos principais tópicos geralmente presentes nesse tipo de documento. Em seguida, solicitou-se ao modelo que produzisse o texto em um formato de anamnese, orientando-o a utilizar expressões como "Sem informações" ou "Não informado" caso não houvesse dados suficientes para preencher algum tópico. Por fim, forneceu-se a transcrição que deveria ser formatada. Um exemplo do prompt pode ser visto a seguir, onde "{ANAMNESE_DEF}" representa a definição da

anamnese e "{transcript}" a transcrição a ser formatada:

Figura 10 – Prompt padrão implementado no modelo phi3

```
<|system|>
Você é um formatador de prontuários e anamneses médicas
em português.
<|end|>
<|user|>
{ANAMNESE_DEF} Formate o seguinte texto no formato de
anamnese, se não houver informações suficientes para um
tópico coloque "Sem informações" ou "Não informado":
<{transcript}>
<|end|>
<|endoftext|>
```

Fonte: Elaborada pela autora.

Observa-se que a formatação do prompt na Figura 10 exibe tags como "<|system|>", "<|user|>" e "<|end|>", além de "<|endoftext|>". Entretanto, a nomenclatura dessas *tags* e a forma de delimitar as mensagens podem variar de acordo com o modelo utilizado.

3.1.2 Modelos Avaliados

Definido o *prompt* a ser utilizado, procedeu-se à aplicação de três modelos de linguagem distintos na tarefa de formatação das transcrições. Desse modo, foi estabelecida uma base de **70 transcrições**, gravadas por alunos da área médica a partir de anamneses reais disponibilizadas pela Hapvida, para ser utilizada como o conjunto de teste nos modelos escolhidos.

O primeiro LLM avaliado foi o Llama-3.2-1B-Instruct¹, que pertence à coleção Llama 3.2 e possui 1 bilhão de parâmetros. Lançado em 25 de setembro de 2024, este modelo é destinado a usos comerciais e de pesquisa em múltiplos idiomas, incluindo aplicações como assistentes de chat, recuperação de conhecimento, sumarização e assistentes de escrita.

O segundo modelo é Mistral-Nemo-Instruct-2407², uma versão ajustada do Mistral-Nemo-2407, desenvolvida em colaboração entre a MistralAI e a NVIDIA. Este modelo possui uma janela de contexto de 128k *tokens* e foi treinado com uma grande proporção de dados multi-lingues e de código, o que o torna altamente eficiente para diversas tarefas de processamento de linguagem natural. Além disso, o Mistral-Nemo-Instruct-2407 funciona como uma substituição

¹ <https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>

² <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>

direta do Mistral 7B, apresentando um desempenho significativamente superior em comparação com modelos de tamanho similar ou menor.

O último modelo avaliado é o Phi-3-mini-4k-instruct³, um modelo de linguagem com 3,8 bilhões de parâmetros, treinado tanto com dados sintéticos quanto com dados filtrados de sites públicos. Ele pertence à família Phi-3, na versão Mini. Ao ser avaliado em *benchmarks* que testam senso comum, compreensão de linguagem, contexto longo, raciocínio lógico e outros tópicos, o modelo demonstrou um desempenho robusto entre os modelos com menos de 13 bilhões de parâmetros.

Em cada um desses três modelos, foi inserido o *prompt* descrito na subseção anterior, garantindo a consistência dos experimentos. Para cada modelo, aplicou-se o *template* correspondente, assegurando que a estrutura de instruções e mensagens fosse processada de forma padronizada.

Ao final dessa etapa, obteve-se três conjuntos de resultados, um para cada LLM avaliado, gerando diferentes versões das anamneses médicas. Essas versões foram, então, submetidas a um processo de avaliação (apresentado na seção 3.4), o qual permitiu comparar aspectos qualitativos e quantitativos das formatações geradas por cada modelo.

3.2 Coleta e Geração de Dados de Treinamento

Dada a ausência de *datasets* específicos para a tarefa de formatação de anamneses médicas, foi necessário criar um conjunto de dados próprio para o treinamento do modelo selecionado no *benchmark* inicial. Assim, extraiu-se um novo conjunto de **100 anamneses** transcritas pelo modelo ASR Whisper⁴ por meio da plataforma da plataforma MedTalkAI (Gonçalves *et al.*, 2024a). Essas transcrições, assim como as utilizadas no conjunto de teste, foram gravadas por três alunos da área médica.

Em seguida, foi criado um conjunto de pares *prompt/resposta*, visando o treinamento descrito na Seção 3.3. As respostas consistem em anamneses formatadas a partir de cada transcrição correspondente, geradas inicialmente pelo modelo GPT o1-mini. Adicionalmente, todas as formatações foram revisadas para assegurar que não houvesse inclusão de informações inexistentes na transcrição original e para confirmar a coerência entre cada tópico e seu respectivo conteúdo.

³ <https://huggingface.co/microsoft/Phi-3-mini-4k-instruct>

⁴ <https://huggingface.co/openai/whisper-large-v3>

Ademais, em cada *prompt*, estabeleceu-se que o LLM deveria se comportar como "médico plantonista, responsável por formatar anamneses médicas". Além disso, incluiu-se na mensagem de definição do comportamento do LLM a definição do que é uma anamnese, bem como sua estrutura geral, reforçando que o modelo "não cometa erros" devido à sensibilidade dos dados. Adicionalmente, estabeleceu-se que, na ausência de informações suficientes para determinado tópico, o LLM deveria ignorar a seção ou preenchê-la com "Sem informações".

Por fim, esses pares de *prompt*/resposta foram ajustados ao *template* esperado pelo modelo a ser treinado (Figura 11), garantindo a padronização necessária para o processo de *fine-tuning*.

Figura 11 – Prompt utilizado para o treinamento

```
<s>
[INST]Você é um médico plantonista, responsável por
formatar anamneses médicas. {ANAMNESE_DEF}. Não cometa
erros, pois os dados que serão trabalhados são
sensíveis. Não insira novas informações. Se algum tópico
não possui informações suficientes ignore ou coloque
'Sem informações'.

Formate o seguinte texto no formato de anamnese:
{transcript}
[/INST]
{answer}
</s>
```

Fonte: Elaborada pela autora.

3.3 Fine-tuning

Após a conclusão do *Benchmark* descrito anteriormente, o modelo Mistral Nemo Instruct foi selecionado para a etapa de *fine-tuning*. O objetivo do treinamento foi aprimorar sua capacidade de formatação de anamneses médicas em português e reduzir as principais limitações identificadas no benchmark.

O treinamento foi realizado em um ambiente com acesso garantido a uma GPU Tesla V100 de 32GB, com disponibilidade com até 30GB de memória RAM e dois núcleos de CPU dedicados. A Tabela 1 apresenta os hiperparâmetros utilizados. O processo foi conduzido por meio das bibliotecas *Transformers* (Wolf *et al.*, 2020) e TRL (*Training Reinforcement Learning*) (Werra *et al.*, 2020), mantendo-se consistência na definição do *prompt* (Seção 3.2).

A estratégia de *fine-tuning* envolveu apenas 1 época de treinamento, com *batch*

Tabela 1 – Hiperparâmetros do treinamento.

Hiperparâmetro	Valor
Número de épocas (<i>epochs</i>)	1
<i>Batch size</i>	4
<i>Gradient accumulation steps</i>	2
<i>Learning rate</i>	2e-4
<i>Weight decay</i>	0.001
<i>fp16</i>	<i>True</i>
<i>Warmup ratio</i>	0.3

Fonte: Elaborado pela autora.

size igual a 4 em cada dispositivo, acumulação de gradiente a cada 2 *steps* e uma taxa de aprendizagem de 2e-4. O *optimizer* configurado em 8 *bits* contribuiu para reduzir a pressão de memória, mantendo a eficiência na atualização dos pesos.

O objetivo principal foi avaliar se seria possível obter maior consistência na formação das anamneses médicas, reduzindo eventuais erros de estruturação e garantindo maior aderência às necessidades clínicas. Ao final do treinamento, obteve-se uma **loss final de 2.6** com um **train runtime de 133.3 segundos**. Por meio dessa etapa, esperava-se melhorar a capacidade do LLM em lidar com detalhes específicos do formato de uma anamnese médica, fornecendo uma base mais sólida para sua aplicação em contextos reais.

Por fim, utilizou-se o modelo ajustado⁵ para gerar novas formatações das transcrições pertencentes ao conjunto de teste. Essas formatações foram, então, submetidas ao mesmo procedimento de avaliação descrito na Seção 3.4, permitindo analisar eventuais melhorias na consistência, clareza e aderência ao formato de anamnese. Dessa forma, foi possível verificar até que ponto o *fine-tuning* contribuiu para aprimorar a capacidade do LLM em lidar com as especificidades da formatação de prontuários e anamneses médicas.

3.4 Avaliação

Para aferir a qualidade das formatações geradas pelos modelos testados, cada anamnese foi avaliada por três alunos da área de saúde: dois graduandos do 10º semestre de Medicina e um graduando do 9º semestre de Enfermagem. Esses avaliadores atribuíram notas em uma escala de 1 a 5, em que 1 representa um resultado ruim e 5, um resultado excelente. Além disso, de forma opcional, eles puderam anotar considerações sobre cada formatação, incluindo a presença de informações inexistentes na transcrição original ou posicionamento incorreto de conteúdo no texto formatado.

⁵ <https://huggingface.co/medtalkai/Mistral-Nemo-Instruct-Anamnese-281224>

Por fim, com base nas pontuações atribuídas, foram calculadas métricas descritivas para cada modelo, incluindo a média das pontuações, os valores mínimo e máximo, além do desvio-padrão. Ademais, para garantir a confiança dos dados, foram aplicadas duas métricas estatísticas: o Coeficiente de Correlação tau de Kendall (Subseção 3.4.1) e o Kappa de Fleiss (Subseção 3.4.2).

3.4.1 Coeficiente de Correlação tau de Kendall

O Coeficiente de Correlação tau de Kendall (Kendall, 1938) é uma medida não-paramétrica que avalia a associação entre duas variáveis por meio de seus postos. Pode ser aplicado tanto a dados contínuos quanto a dados ordinais, e é especialmente adequado em situações em que não se pode assumir uma distribuição normal. O valor de τ pode ser obtido de acordo com a Equação (3.1):

$$\tau = \frac{c - d}{c + d} = \frac{S}{\binom{n}{2}} = \frac{2S}{n(n-1)} \quad (3.1)$$

onde c é o número de pares concordantes e d é o número de pares discordantes; S representa a soma total resultante da comparação de postos. Caso existam empates entre as variáveis, emprega-se a Equação (3.2):

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T} \sqrt{n(n-1)/2 - U}} \quad (3.2)$$

$$T = \sum_t t(t-1)/2$$

$$U = \sum_u u(u-1)/2$$

onde t corresponde ao número de observações empatadas na variável x e u ao número de observações amarradas na variável y . Os valores de τ variam de -1 a $+1$: magnitudes próximas a ± 1 indicam correlações fortes, enquanto valores próximos de 0 sugerem correlações fracas ou inexistentes. A implementação utilizada para o cálculo do tau de Kendall dentro deste trabalho foi realizada utilizando a biblioteca Python **SciPy** (Virtanen *et al.*, 2020).

3.4.2 Kappa de Fleiss

O Kappa de Fleiss (Fleiss, 1971) é derivado do Kappa de Cohen (Cohen, 1960), sendo a principal diferença entre as duas o número de avaliadores, onde o de Cohen só é aplicado para cenários com apenas 2 observadores e o de Fleiss pode ser utilizado em cenário com 3 ou mais.

Assim como o Coeficiente de Correlação de Kendall, os valores possíveis variam de -1 a $+1$, porém, nesse caso, os valores próximos a -1 indicam pouca concordância entre os avaliadores e valores próximos a $+1$ indicam uma concordância perfeita. Complementando, Fleiss (1971) define que um coeficiente Kappa entre 0,40 e 0,60 indica uma concordância regular, entre 0,61 e 0,75 indica uma concordância boa e acima de 0,75 indica uma concordância excelente.

A definição matemática do Kappa de Fleiss é baseada em uma matriz de observações $N \times k$, onde n_{ij} representa o número de avaliadores que atribuíram a i -ésima observação à j -ésima categoria. As seguintes equações são utilizadas para seu cálculo:

$$\begin{aligned}
 p_j &= \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \\
 P_i &= \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1) = \frac{1}{n(n-1)} \left(\sum_{j=1}^k n_{ij}^2 - n \right), \\
 P_o &= \frac{1}{N} \sum_{i=1}^N P_i, \\
 P_e &= \sum_{j=1}^k p_j^2,
 \end{aligned} \tag{3.3}$$

onde P_j é a proporção de todas as atribuições à j -ésima categoria, P_i é o grau de concordância entre os n avaliadores para a i -ésima observação, P_o é a concordância observada geral, e P_e é a proporção esperada de concordância devido ao acaso. O Kappa de Fleiss é então definido pela Equação (3.4):

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{3.4}$$

A implementação do Kappa de Fleiss utilizada nas avaliações foi realizada através da biblioteca Python **statsmodels** (Seabold; Perktold, 2010).

4 RESULTADOS

Neste capítulo, apresentam-se os resultados alcançados durante o desenvolvimento deste trabalho. O capítulo está organizado em duas seções: a primeira apresenta os resultados obtidos no *Benchmark* inicial dos modelos (Seção 4.1), enquanto a segunda descreve os resultados alcançados durante o processo de *Fine-tuning* para a tarefa de formatação de anamneses médicas (Seção 4.2). Todos os exemplos de anamneses formatadas são baseados na mesma transcrição (Figura 12).

Figura 12 – Exemplo de transcrição utilizada durante o processo de avaliação

Paciente retorna com exame sem alterações. Ultrassom axilar indica presença de cisto. Nega queixas. Alergias medicamentosas, nega. Patologias prévias, nega. Medicamento em uso, nega. Ao exame físico, bom estado geral, lúcida e orientada em tempo e espaço. Padrão respiratório confortável em ar ambiente. Sem sinais de desidratação. Aparelho pulmonar, murmúrio vesicular universal sem ruídos adventícios Aparelho cardiovascular Ritmo cardíaco regular em dois tempos, sem sopros Abdomen, flácido, ruídos hidroaéreos presentes e dolor à palpação Sem sinais de irritação peritoneal Giordano negativo Neuro, vigio, pupilas isocóricas e fotorreagentes sem déficits focais, sem sinais de irritação manígia, força muscular simétrica e preservada em membros, sem desvio de comissura labial, extremidades bem perfundidas, acianóticas sem edemas, sem sinais de empastamento em panturrilhas, hoste muscular, ausência de sinais fosíticos ou limitação da movimentação em articulações.

Fonte: Elaborada pela autora.

4.1 Resultados do *Benchmark*

Esta seção apresenta os resultados obtidos a partir da avaliação de desempenho de três modelos: Phi3, Llama e Mistral. O objetivo deste *benchmark* foi avaliar a capacidade dos modelos em gerar respostas consistentes e alinhadas aos critérios estabelecidos, antes da etapa de *fine-tuning*. A análise inclui as opiniões de três avaliadores independentes e fornece métricas descritivas, coeficientes de concordância e correlação para cada modelo.

4.1.1 Modelo Phi3

Os resultados gerais do modelo Phi3 indicam um desempenho moderado, com uma média de 3.22, mediana de 3, moda de 3 e desvio padrão de 0.96, evidenciando baixa variabilidade entre as avaliações. Na Tabela 2, são apresentadas as métricas descritivas por avaliador. Observa-se que as médias variaram entre 2.94 e 3.39, com consistência nos desvios

padrão, que oscilaram entre 0.87 e 0.99.

Tabela 2 – Resultados descritivos do modelo Phi3

Métrica	Avaliador 01	Avaliador 02	Avaliador 03
Média	2.94	3.29	3.39
Mediana	3	3	3
Moda	3	3	3
Desvio Padrão	0.99	0.97	0.88
Mínimo	1	1	1
Máximo	5	5	5

Fonte: Elaborado pela autora.

O coeficiente Kappa entre os avaliadores foi de 0.179, indicando um nível de concordância baixo. Além disso, a Tabela 3 apresenta os coeficientes de correlação de Kendall entre os pares de avaliadores, que revelam níveis moderados de correlação, com valores variando de 0.29 a 0.52.

Tabela 3 – Coeficiente de correlação *tau* de Kendall para o modelo Phi3

Par de Avaliadores	Coeficiente τ	<i>p-value</i>
Avaliador 01 & Avaliador 02	0.52	3.51e-07
Avaliador 01 & Avaliador 03	0.35	6.56e-04
Avaliador 02 & Avaliador 03	0.29	4.62e-03

Fonte: Elaborado pela autora.

Por fim, os avaliadores apontaram críticas específicas para cada formatação. As principais críticas ao modelo Phi3 foram relacionadas à falta de precisão nas respostas, com informações parcialmente corretas ou incompletas. Além disso, foi mencionada a utilização excessiva de linguagem genérica, o que dificultaria a aplicação prática das soluções geradas. Um exemplo de formatação realizada pelo Phi3 pode ser encontrado na Figura 13.

4.1.2 Modelo Llama

Os resultados gerais do modelo Llama indicam um desempenho inferior ao do Phi3, com uma média de 2.01, mediana de 2, moda de 1 e desvio padrão de 1.03. Esses valores sugerem maior variabilidade e avaliações menos consistentes entre os avaliadores. Na Tabela 4, estão detalhadas as métricas por avaliador, com médias variando de 1.57 a 2.65.

O coeficiente Kappa para o modelo Llama foi de 0.093, reforçando a baixa concordância entre os avaliadores. Já os coeficientes de Kendall, apresentados na Tabela 5, destacam correlações significativas, mas ainda inferiores às observadas no modelo Phi3.

Figura 13 – Exemplo de formatação feita pelo Phi3

Identificação:
- Nome: Sem informações
- Idade: Sem informações
- Sexo: Sem informações
- Estado civil: Sem informações
- Profissão: Sem informações
- Naturalidade: Sem informações
Queixa principal:
- Sem queixa
História da doença atual:
- Sem informações
História patológica pregressa:
- Sem informações
História psicossocial:
- Sem informações
História familiar:
- Sem informações
Exame físico:
- Estado geral: Bom
- Padrão respiratório: Confortável em ar ambiente
- Sinais de desidratação: Não presentes
- Aparelho pulmonar: Murmúrio vesicular universal sem ruídos adventícios
- Aparelho cardiovascular: Ritmo cardíaco regular em dois tempos, sem sopros
- Abdome: Flácido, ruídos hidroaéreos presentes e dolor à palpação
- Sinais de irritação peritoneal: Não presentes
- Neuro: Sem déficits focais
- Vício: Sem sinais de irritação manígia
- Pupilas: Isocóricas e fotorreagentes
- Força muscular: Simétrica e preservada em membros
- Comissura labial: Não presentes
- Extremidades: Acianóticas sem edemas
- Sinais de empastamento: Não presentes
- Hoste muscular: Não presentes
- Movimentação articular: Não presentes
IOA (Interrogatório por Órgãos e Aparelhos):
- Sem informações
Exames complementares:
- Ultrassom axilar: Presença de cisto
Hipótese diagnóstica:
- Sem informações
Conduta:
- Sem informações

Fonte: Elaborada pela autora.

Tabela 4 – Resultados descritivos do modelo Llama

Métrica	Avaliador 01	Avaliador 02	Avaliador 03
Média	1.57	1.77	2.65
Mediana	1	2	3
Moda	1	1	3
Desvio Padrão	0.72	0.93	1.03
Mínimo	1	1	1
Máximo	3	5	4

Fonte: Elaborado pela autora.

O modelo Llama enfrentou severas críticas por fornecer respostas frequentemente fora do contexto ou irrelevantes em relação ao problema apresentado. Também foram feitas reclamações sobre a inconsistência no formato das respostas, com variações que dificultavam a

Tabela 5 – Coeficiente de correlação *tau* de Kendall para o modelo Llama

Par de Avaliadores	Coeficiente τ	<i>p-value</i>
Avaliador 01 & Avaliador 02	0.28	9.69e-03
Avaliador 01 & Avaliador 03	0.55	2.76e-07
Avaliador 02 & Avaliador 03	0.20	5.90e-02

Fonte: Elaborado pela autora.

leitura e a comparação dos resultados. Um exemplo de formatação realizada pelo Llama pode ser encontrado na Figura 14.

4.1.3 Modelo Mistral

O modelo Mistral apresentou o melhor desempenho geral no *benchmark*, com uma média de 4.00, mediana de 4, moda de 5 e desvio padrão de 0.93. Esses resultados evidenciam avaliações consistentemente altas. Na Tabela 6, estão descritas as métricas descritivas detalhadas por avaliador, com médias variando de 3.84 a 4.30 e baixos desvios padrão, o que indica menor variabilidade.

Tabela 6 – Resultados descritivos do modelo Mistral

Métrica	Avaliador 01	Avaliador 02	Avaliador 03
Média	3.84	3.86	4.30
Mediana	4	4	4
Moda	3	5	5
Desvio Padrão	0.96	0.99	0.77
Mínimo	2	2	3
Máximo	5	5	5

Fonte: Elaborado pela autora.

O coeficiente Kappa foi de 0.266, o maior entre os modelos avaliados, refletindo uma concordância moderada. Já a Tabela 7 apresenta os coeficientes de Kendall, que indicam correlações similares às observadas nos outros modelos.

Tabela 7 – Coeficiente de correlação *tau* de Kendall para o modelo Mistral

Par de Avaliadores	Coeficiente τ	<i>p-value</i>
Avaliador 01 & Avaliador 02	0.59	1.99e-08
Avaliador 01 & Avaliador 03	0.30	4.92e-03
Avaliador 02 & Avaliador 03	0.27	1.24e-02

Fonte: Elaborado pela autora.

Embora o modelo Mistral tenha obtido os melhores resultados quantitativos, algumas

Figura 14 – Exemplo de formatação feita pelo Llama

```

**Anamnese Completa do Paciente**
**Identificação**
* Nome: [Não informado]
* Idade: [Não informado]
* Sexo: [Não informado]
* Estado Civil: [Não informado]
* Profissão: [Não informado]
* Naturalidade: [Não informado]
**Queixa Principal**
A queixa principal é a presença de um cisto no ultrassom axilar.
**História da Doença Atual**
A história da doença atual é que a paciente tem apresentado
sintomas de cisto no ultrassom axilar, que foram diagnosticados
durante a avaliação física.
**História Patológica Progredida**
A história patológica progredida é que a paciente não apresenta
nenhuma condição prévia.
**História Psicossocial**
A história psicossocial é que a paciente tem um estilo de vida
saudável e não apresenta hábitos ou questões que possam influenciar
sua saúde.
**História Familiar**
A história familiar é que a paciente não apresenta nenhuma doença
crônica ou genética prévia.
**Exame Físico**
A avaliação física foi realizada e os seguintes sinais foram
observados:
* Padrão respiratório confortável em ar ambiente.
* Sinais de desidratação não foram observados.
* Aparelho pulmonar, murmúrio vesicular universal sem ruídos
adventícios.
* Abdomen, flácido, ruídos hidroaéreos presentes e dor à
palpação.
* Sinais de irritação peritoneal negativos.
**IOA (Interrogatório por Órgãos e Aparelhos)**
O interrogatório por órgãos e aparelhos foi realizado e os
seguintes sinais foram observados:
* Ritmo cardíaco regular em dois tempos, sem sopros.
* Abdomen, flácido, ruídos hidroaéreos presentes e dor à
palpação.
* Sinais de irritação peritoneal negativos.
**Exames Complementares**
A anamnese foi acompanhada por exames complementares, incluindo
exames de ultrassom, ressonância magnética, etc.
**Hipótese Diagnóstica**
A hipótese diagnóstica é que a paciente apresenta um cisto no
ultrassom axilar, que foi diagnosticado durante a avaliação física.
**Conduta**
A conduta recomendada é a realização de exames adicionais, como
exames de ultrassom, ressonância magnética, etc., para confirmar ou
descartar suspeitas diagnósticas, além de realizar tratamentos e
intervenções adequadas para o cisto.

```

Fonte: Elaborada pela autora.

críticas pontuais foram registradas. A principal delas foi a sobrecarga de detalhes em algumas respostas, o que, embora demonstrasse profundidade, as tornava menos objetivas e, por vezes, mais difíceis de interpretar. Um exemplo de formatação gerada pelo Mistral pode ser encontrado na Figura 15.

Figura 15 – Exemplo de formatação feita pelo Mistral

```

**ANÁMNESE**
**Identificação:**
- Nome: Não informado
- Idade: Não informado
- Sexo: Não informado
- Estado civil: Não informado
- Profissão: Não informado
- Naturalidade: Não informado
**Queixa principal:**
- Paciente retorna com exame sem alterações. Ultrassom axilar
indica presença de cisto.
**História da doença atual:**
- Sem informações
**História patológica pregressa:**
- Nega patologias prévias.
**História psicossocial:**
- Sem informações
**História familiar:**
- Sem informações
**Exame físico:**
- Bom estado geral, lúcida e orientada em tempo e espaço.
- Padrão respiratório confortável em ar ambiente.
- Sem sinais de desidratação.
- Aparelho pulmonar: murmúrio vesicular universal sem ruídos
adventícios
- Aparelho cardiovascular: Ritmo cardíaco regular em dois tempos,
sem sopros
- Abdomen: Flácido, ruídos hidroaéreos presentes e dolor à
palpação
- Neuro: Vívido, pupilas isocóricas e fotorreagentes sem déficits
focais, sem sinais de irritação meningiana, força muscular
simétrica e preservada em membros, sem desvio de comissura labial,
extremidades bem perfundidas, acianóticas sem edemas, sem sinais de
empastamento em panturrilhas, hoste muscular, ausência de sinais
fosíuticos ou limitação da movimentação em articulações.
**IOA (Interrogatório por Órgãos e Aparelhos):**
- Sem informações
**Exames complementares:**
- Ultrassom axilar: Indica presença de cisto
**Hipótese diagnóstica:**
- Sem informações
**Conduta:**
- Sem informações

```

Fonte: Elaborada pela autora.

4.2 Resultados do *Fine-tuning*

Com base nos resultados obtidos no *benchmark*, o modelo Mistral passou por *fine-tuning*, utilizando um conjunto de dados distinto daquele empregado na avaliação inicial. O objetivo foi aprimorar a precisão e a consistência das respostas, levando em consideração tanto os *feedbacks* qualitativos quanto os quantitativos fornecidos pelos avaliadores

Com o modelo ajustado, foram geradas novas formatações para os mesmos dados utilizados no *benchmark*. No entanto, o modelo não apresentou melhoria geral, registrando uma média de 3.89, ligeiramente inferior à média do modelo original, que era 4.00. Por outro lado, os valores de mediana (4) e moda (5) permaneceram inalterados. A Tabela 8 apresenta as métricas

descritivas das avaliações realizadas por cada avaliador para o modelo ajustado após o processo de *fine-tuning*

Tabela 8 – Resultados Descritivos do modelo Mistral com Fine-tuning

Métrica	Avaliador 01	Avaliador 02	Avaliador 03
Média	4.014493	3.811594	3.884058
Mediana	4	4	4
Moda	5	5	4
Desvio Padrão	0.977583	1.239964	1.007855
Mínimo	1	1	1
Máximo	5	5	5

Fonte: Elaborada pelo autora.

Apesar da leve queda na média geral, o coeficiente Kappa aumentou para 0.387, indicando uma concordância moderada entre os avaliadores, em comparação com o modelo original. Os coeficientes de correlação de Kendall, apresentados na Tabela 9, também sugerem uma maior consistência nas avaliações realizadas pelos diferentes avaliadores, com coeficientes variando de 0.45 a 0.68. Um exemplo de formatação feita por este modelo pode ser encontrado na Figura 16.

Tabela 9 – Coeficiente de Correlação tau de Kendall entre avaliadores para o modelo Mistral com Fine-tuning

Par de Avaliadores	Coeficiente τ	p-value
Avaliador 01 & Avaliador 02	0.782945	1.990200e-08
Avaliador 01 & Avaliador 03	0.309200	3.070804e-03
Avaliador 02 & Avaliador 03	0.320354	1.816091e-03

Fonte: Elaborada pelo autora.

Embora o modelo ajustado tenha obtido uma leve melhora na concordância entre os avaliadores, conforme evidenciado pelos coeficientes Kappa e Kendall, a análise qualitativa demonstra que as alterações realizadas durante o *fine-tuning* não resultaram em uma melhora perceptível na experiência geral dos avaliadores. Foi destacado que, em alguns casos, as formatações se tornaram excessivamente detalhadas, comprometendo a objetividade, que é essencial em uma anamnese. Além disso, apesar de o modelo ajustado ter gerado melhorias em algumas respostas específicas, ele também apresentou novos problemas, como redundância em certas formatações e dificuldades na interpretação de elementos contextuais complexos. Essa piora no desempenho do modelo ajustado pode ser atribuída a dois fatores principais.

Primeiramente, destaca-se que o treinamento realizado foi muito curto, com apenas

Figura 16 – Exemplo de formatação feita pelo Mistral com Fine-tuning

Identificação: Sem informações.
 Queixa principal: Retorno com exame sem alterações.
 História da doença atual: Nega queixas.
 História patológica pregressa: Nega patologias prévias.
 História psicossocial: Sem informações.
 História familiar: Sem informações.
 Exame físico: Bom estado geral, lúcida e orientada em tempo e espaço. Padrão respiratório confortável em ar ambiente. Sem sinais de desidratação. Aparelho pulmonar, murmúrio vesicular universal sem ruídos adventícios. Aparelho cardiovascular, ritmo cardíaco regular em dois tempos, sem sopros. Abdômen, flácido, ruídos hidroaéreos presentes e dolor à palpação. Sem sinais de irritação peritoneal. Giordano negativo. Neuro, vígil, pupilas isocóricas e fotorreagentes sem déficits focais, sem sinais de irritação meníngea, força muscular simétrica e preservada em membros, sem desvio de comissura labial, extremidades bem perfundidas, acianóticas sem edemas, sem sinais de empastamento em panturrilhas, hoste muscular, ausência de sinais fosícticos ou limitação da movimentação em articulações.
 IOA (Interrogatório por Órgãos e Aparelhos): Sem informações.
 Exames complementares: Ultrassom axilar indica presença de cisto.
 Hipótese diagnóstica: Sem informações.
 Conduta: Sem informações.

Fonte: Elaborada pela autora.

uma época. Esse período reduzido pode ter sido insuficiente para que o modelo ajustado aprendesse padrões relevantes ou refinasse as capacidades previamente existentes. Além disso, ao final do treinamento, o valor da *loss* permaneceu elevado, o que sugere que o modelo não convergiu adequadamente ou que os ajustes realizados não foram suficientes para reduzir os erros nas predições. Um treinamento mais longo poderia ter favorecido melhores resultados.

Outro fator a ser considerado é a qualidade do conjunto de dados utilizado no *fine-tuning*. Os dados de treinamento foram gerados por outro LLM e não passaram por um processo de correção ou validação por especialistas da área médica. A ausência de supervisão especializada pode ter introduzido padrões inadequados ou inconsistências no treinamento, o que levou o modelo ajustado a apresentar limitações adicionais, como redundâncias ou dificuldades em interpretar contextos mais complexos.

A soma desses dois fatores — treinamento curto e dados de baixa qualidade — pode ter resultado em um modelo ajustado que, além de não melhorar a precisão, apresentou novas limitações. Esses resultados destacam a necessidade de investir em um treinamento mais robusto, utilizando um conjunto de dados revisado e validado por especialistas, além de garantir uma maior atenção ao ajuste dos hiperparâmetros e ao monitoramento das métricas durante o treinamento.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo explorar o desempenho de diferentes modelos de linguagem na tarefa de formatação de anamneses médicas, com foco no uso de técnicas de *fine-tuning* para aprimorar a consistência e a precisão das respostas geradas. O estudo foi realizado em duas etapas principais: a primeira consistiu em um *benchmark* inicial dos modelos Phi3, Llama e Mistral, enquanto a segunda envolveu a aplicação de *fine-tuning* no modelo Mistral, a fim de aprimorar suas respostas.

Os resultados do *benchmark* indicaram que o modelo Mistral apresentou o melhor desempenho entre os modelos avaliados, com notas consistentemente altas e uma maior concordância entre os avaliadores, refletida pelos coeficientes de Kendall e Fleiss. No entanto, embora o modelo tenha se destacado quantitativamente, algumas críticas foram registradas quanto à sobrecarga de detalhes nas respostas, que dificultavam a interpretação e tornavam as respostas menos objetivas.

O modelo Llama apresentou o pior desempenho, com avaliações muito discrepantes entre os avaliadores e respostas frequentemente fora do contexto ou irrelevantes. O modelo Phi3, apesar de um desempenho moderado, apresentou falhas como falta de precisão e uso excessivo de linguagem genérica.

O *fine-tuning* realizado no modelo Mistral, baseado nas críticas qualitativas e quantitativas dos avaliadores, não resultou em uma melhoria significativa nos resultados.. A média de avaliação do modelo ajustado foi ligeiramente inferior à do modelo original, embora os valores de mediana e moda tenham se mantido os mesmos. Isso sugere que o *fine-tuning* não foi capaz de resolver as críticas sobre a sobrecarga de detalhes nas respostas.

As principais contribuições deste trabalho incluem a análise comparativa dos modelos de linguagem na formatação de anamneses médicas e a aplicação do *fine-tuning* como técnica de aprimoramento. Embora o modelo Mistral tenha sido o mais promissor, as críticas apontam áreas para melhoria, especialmente quanto à objetividade e clareza das respostas.

Embora os resultados obtidos sejam promissores, a pesquisa apresenta algumas limitações que devem ser consideradas. A principal limitação diz respeito à qualidade dos dados utilizados, que pode impactar a precisão e a generalização dos resultados. Além disso, um possível viés entre os avaliadores e os modelos avaliados pode ter influenciado os resultados. Como as avaliações foram realizadas de forma sequencial, a avaliação de um modelo pode ter sido influenciada pela avaliação anterior, o que pode ter gerado um efeito de contaminação nos

resultados das avaliações subsequentes.

Como trabalhos futuros, pretende-se realizar um treinamento mais longo do modelo, o que pode melhorar sua capacidade de generalização e sua acurácia em contextos diversos. Além disso, pretende-se validar e expandir o conjunto de treinamento, incorporando um maior número de casos clínicos e abordagens diversificadas. A exploração de novas técnicas de aprendizado de máquina, como redes neurais profundas, estratégias de *prompt* e ajustes de hiperparâmetros, também são direções importantes para a melhoria contínua do modelo.

REFERÊNCIAS

- ABDIN, M.; JACOBS, S. A.; AWAN, A. A.; ANEJA, J.; AWADALLAH, A.; AWADALLA, H.; BACH, N.; BAHREE, A.; BAKHTIARI, A.; BEHL, H. *et al.* **Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone**. 2024. Disponível em: <https://arxiv.org/abs/2404.14219>.
- BAEVSKI, A.; ZHOU, Y.; MOHAMED, A.; AULI, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 12449–12460. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.
- BALAJI, V.; SADASHIVAPPA, G. Speech disabilities in adults and the suitable speech recognition software tools-a review. In: IEEE. **2015 International Conference on Computing and Network Communications (CoCoNet)**. [S. l.], 2015. p. 559–564.
- BENSEÑOR, I. M. Anamnese, exame clínico e exames complementares como testes diagnósticos. **Revista de Medicina**, v. 92, n. 4, p. 236–241, dez. 2013. Disponível em: <https://www.revistas.usp.br/revistadc/article/view/85896>.
- CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. BPLN, 2024. ISBN 978-65-00-95750-1. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/>.
- CHURCH, K. W.; CHEN, Z.; MA, Y. Emerging trends: A gentle introduction to fine-tuning. **Natural Language Engineering**, v. 27, n. 6, p. 763–778, 2021.
- COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37 – 46, 1960.
- DALIANIS, H. **Clinical text mining: Secondary use of electronic patient records**. [S. l.]: Springer Nature, 2018.
- FELIX, V. G.; MENA, L. J.; OSTOS, R.; MAESTRE, G. E. A pilot study of the use of emerging computer technologies to improve the effectiveness of reading and writing therapies in children with down syndrome. **British Journal of Educational Technology**, Wiley Online Library, v. 48, n. 2, p. 611–624, 2017.
- FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological bulletin**, American Psychological Association, v. 76, n. 5, p. 378, 1971.
- FUEHRER, S.; WEIL, A.; OSTERBERG, L. G.; ZULMAN, D. M.; MEUNIER, M. R.; SCHWARTZ, R. Building authentic connection in the patient-physician relationship. **Journal of Primary Care & Community Health**, v. 15, p. 21501319231225996, 2024. PMID: 38281122. Disponível em: <https://doi.org/10.1177/21501319231225996>.
- GARRETT, L. J.; HAMMOND, W.; STEAD, W. The effects of computerized medical records on provider efficiency and quality of care. **Methods of information in medicine**, Schattauer GmbH, v. 25, n. 03, p. 151–157, 1986.

GONÇALVES, Y.; ALVES, J.; SÁ, B.; SILVA, L.; MACEDO, J.; SILVA, T. C. da. MedTalkAI: Assisted anamnesis creation with automatic speech recognition. In: **Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados**. Porto Alegre, RS, Brasil: SBC, 2024. p. 83–88. ISSN 0000-0000. Disponível em: https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/30775.

GONÇALVES, Y.; ALVES, J.; SÁ, B.; SILVA, L.; MACEDO, J.; SILVA, T. C. da. Speech recognition models in assisting medical history. In: **Anais do XXXIX Simpósio Brasileiro de Bancos de Dados**. Porto Alegre, RS, Brasil: SBC, 2024. p. 485–497. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/30715>.

GÜR, B. **Improving speech recognition accuracy for clinical conversations**. Tese (Doutorado) – Massachusetts Institute of Technology, 2012.

HADI, M. U.; QURESHI, R.; SHAH, A.; IRFAN, M.; ZAFAR, A.; SHAIKH, M. B.; AKHTAR, N.; WU, J.; MIRJALILI, S. *et al.* Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. **Authorea Preprints**, Authorea, 2023.

JIANG, A. Q.; SABLAYROLLES, A.; MENSCH, A.; BAMFORD, C.; CHAPLOT, D. S.; CASAS, D. de las; BRESSAND, F.; LENGYEL, G.; LAMPLE, G.; SAULNIER, L.; LAVAUD, L. R.; LACHAUX, M.-A.; STOCK, P.; SCAO, T. L.; LAVRIL, T.; WANG, T.; LACROIX, T.; SAYED, W. E. **Mistral 7B**. 2023.

JIN, D.; JIN, Z.; HU, Z.; VECHTOMOVA, O.; MIHALCEA, R. Deep learning for text style transfer: A survey. **Computational Linguistics**, v. 48, n. 1, p. 155–205, 04 2022. ISSN 0891-2017. Disponível em: https://doi.org/10.1162/coli_a_00426.

KARPAGAVALLI, S.; CHANDRA, E. A review on automatic speech recognition architecture and approaches. **International Journal of Signal Processing, Image Processing and Pattern Recognition**, v. 9, n. 4, p. 393–404, 2016.

KENDALL, M. G. A new measure of rank correlation. **Biometrika**, Oxford University Press, v. 30, n. 1-2, p. 81–93, 1938.

KODISH-WACHS, J.; AGASSI, E.; III, P. K.; OVERHAGE, J. M. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. **AMIA Annual Symposium Proceedings**. [S. l.], 2018. v. 2018, p. 683.

LAI, W.; HANGYA, V.; FRASER, A. Style-specific neurons for steering llms in text style transfer. In: **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**. [S. l.: s. n.], 2024. p. 13427–13443.

LEE, J.; YOON, W.; KIM, S.; KIM, D.; KIM, S.; SO, C. H.; KANG, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, v. 36, n. 4, p. 1234–1240, 09 2019. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btz682>.

LIU, F.; TUR, G.; HAKKANI-TÜR, D.; YU, H. Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 625–630, 2011.

- LUO, R.; SUN, L.; XIA, Y.; QIN, T.; ZHANG, S.; POON, H.; LIU, T.-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. **Briefings in Bioinformatics**, v. 23, n. 6, p. bbac409, 09 2022. ISSN 1477-4054. Disponível em: <https://doi.org/10.1093/bib/bbac409>.
- MACHADO, L. J. C.; RIBEIRO, M. M. F.; GUIMARÃES, M. M. M. Entrevista e anamnese: uma análise crítica dos vários autores e uma proposta de roteiro e de registro de consulta baseados no método clínico centrado no paciente. **Revista Médica de Minas Gerais**, Universidade Federal de Minas Gerais, 2018.
- MALIK, M.; MALIK, M. K.; MEHMOOD, K.; MAKHDOOM, I. Automatic speech recognition: a survey. **Multimedia Tools and Applications**, v. 80, n. 6, p. 9411–9457, Mar 2021. ISSN 1573-7721. Disponível em: <https://doi.org/10.1007/s11042-020-10073-7>.
- MATA, B. U.; N, R. G.; KUMAR, N.; B, R. Effects of computerized patient records on patients. **Bonfring International Journal of Man Machine Interface**, v. 3, p. 25–29, 07 2015.
- MCDONALD, D. D.; PUSTEJOVSKY, J. D. A computational theory of prose style for natural language generation. In: **Proceedings of the Second Conference on European Chapter of the Association for Computational Linguistics**. USA: Association for Computational Linguistics, 1985. (EACL '85), p. 187–193. Disponível em: <https://doi.org/10.3115/976931.976958>.
- MUKHERJEE, S.; DUŠEK, O. **Text Style Transfer: An Introductory Overview**. 2024. Disponível em: <https://arxiv.org/abs/2407.14822>.
- MUKHERJEE, S.; OJHA, A. K.; DUŠEK, O. Are large language models actually good at text style transfer? **arXiv preprint arXiv:2406.05885**, 2024.
- NAVEED, H.; KHAN, A. U.; QIU, S.; SAQIB, M.; ANWAR, S.; USMAN, M.; AKHTAR, N.; BARNES, N.; MIAN, A. **A Comprehensive Overview of Large Language Models**. 2024.
- NGUEAJIO, M. K.; WASHINGTON, G. Hey asr system! why aren't you more inclusive? In: CHEN, J. Y. C.; FRAGOMENI, G.; DEGEN, H.; NTOA, S. (Ed.). **HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence**. Cham: Springer Nature Switzerland, 2022. p. 421–440. ISBN 978-3-031-21707-4.
- O'SHAUGHNESSY, D. Invited paper: Automatic speech recognition: History, methods and challenges. **Pattern Recognition**, v. 41, n. 10, p. 2965–2979, 2008. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320308001799>.
- PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: A method for automatic evaluation of machine translation. In: **Proceedings of the 40th ACL**. USA: Association for Computational Linguistics, 2002. p. 311–318. Disponível em: <https://doi.org/10.3115/1073083.1073135>.
- RADFORD, A.; KIM, J. W.; XU, T.; BROCKMAN, G.; MCLEAVEY, C.; SUTSKEVER, I. Robust speech recognition via large-scale weak supervision. In: PMLR. **International conference on machine learning**. [S. l.], 2023. p. 28492–28518.
- REDDY, D. Speech recognition by machine: A review. **Proceedings of the IEEE**, v. 64, n. 4, p. 501–531, 1976.

RIEMANN, M.; KNIPFER, C.; ROHDE, M.; ADLER, W.; SCHUSTER, M.; NOETH, E.; OETTER, N.; SHAMS, N.; NEUKAM, F.-W.; STELZLE, F. Oral squamous cell carcinoma of the tongue: Prospective and objective speech evaluation of patients undergoing surgical therapy. **Head & Neck**, Wiley Online Library, v. 38, n. 7, p. 993–1001, 2016.

SAPIR, E. **Language; an introduction to the study of speech**. Oxford, England: Harcourt, Brace, 1921. (Language; an introduction to the study of speech.).

SEABOLD, S.; PERKTOLD, J. statsmodels: Econometric and statistical modeling with python. In: **9th Python in Science Conference**. [S. l.: s. n.], 2010.

SOARES, M. O. M.; HIGA, E. d. F. R.; GOMES, L. F.; MARVÃ, J. P. Q.; GOMES, A. I. da F.; GONÇALVES, A. H. C. Impacto da anamnese para o cuidado integral: visão dos estudantes portugueses. **Revista Brasileira em Promoção da Saúde**, v. 29, p. 66–75, 2016.

Toledo, T. F. de; LEE, H. D.; SPOLAÔR, N.; Rodrigues Coy, C. S.; WU, F. C. Web system prototype based on speech recognition to construct medical reports in brazilian portuguese. **International Journal of Medical Informatics**, v. 121, p. 39–52, 2019. ISSN 1386-5056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1386505618302879>.

VASQUEZ-CORREA, J.; OROZCO-ARROYAVE, J.; NÖTH, E. Word accuracy and dynamic time warping to assess intelligibility deficits in patients with parkinsons disease. In: IEEE. **2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)**. [S. l.], 2016. p. 1–5.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: **NIPS**. [S. l.: s. n.], 2017. p. 6000–6010. ISBN 9781510860964.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; van der Walt, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VanderPlas, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020.

WERRA, L. von; BELKADA, Y.; TUNSTALL, L.; BEECHING, E.; THRUSH, T.; LAMBERT, N.; HUANG, S.; RASUL, K.; GALLOUÉDEC, Q. **TRL: Transformer Reinforcement Learning**. [S. l.]: GitHub, 2020. <https://github.com/huggingface/trl>.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; SCAO, T. L.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. M. Transformers: State-of-the-art natural language processing. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**. Online: Association for Computational Linguistics, 2020. p. 38–45. Disponível em: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

YEHIA, A. C.; VIANA, P. R. L.; MACEDO, M. V. M.; DIAS, N. C. de S.; CAMPOS, C. C.; JARDIM, S. N.; GARCIA, J. N. A. de A. Anamnese na prática clínica: uma revisão sobre suas

aplicações e importância. **Revista da Sociedade Brasileira de Clínica Médica**, v. 22, n. 2, p. 116–120, 2024.

YOSHIKAWA, G.; CASTRO, R. C. **Manual de semiologia médica: a prática do exame físico**. 2. ed. [S. l.]: EDUEPA, 2015. ISBN 978-85-8458-005-7.

YUAN, Z.; LIU, J.; ZI, Q.; LIU, M.; PENG, X.; LOU, Y. Evaluating instruction-tuned large language models on code comprehension and generation. **arXiv preprint arXiv:2308.01240**, 2023.