ANNA PAULA FIGUEIREDO GONÇALVES

AVALIAÇÃO DE MODELOS DE LINGUAGEM PARA CRIAÇÃO DE *CHATBOTS* PARA A COMUNICAÇÃO COM PESSOAS SURDAS QUE USAM O PORTUGUÊS COMO SEGUNDA LÍNGUA

LAVRAS – MG

2025

ANNA PAULA FIGUEIREDO GONÇALVES

AVALIAÇÃO DE MODELOS DE LINGUAGEM PARA CRIAÇÃO DE *CHATBOTS* PARA A COMUNICAÇÃO COM PESSOAS SURDAS QUE USAM O PORTUGUÊS COMO SEGUNDA LÍNGUA

Dissertação de Mestrado apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração Inteligência Artificial e Otimização, para a obtenção do título de Mestra.

Prof. Dr. Denilson Alves Pereira
Orientador

Prof. Dr. André Pimenta Freire
Co-orientador

LAVRAS – MG
2025

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Gonçalves, Anna Paula Figueiredo

Avaliação de Modelos de Linguagem para Criação de *Chatbots* para a Comunicação com Pessoas Surdas que usam o Português como Segunda Língua / Anna Paula Figueiredo Gonçalves, Denilson Alves Pereira, André Pimenta Freire. 6^a ed. rev., atual. e ampl. – Lavras : UFLA, 2025.

140p. : il.

Dissertação(mestrado)–Universidade Federal de Lavras, 2025.

Orientador: Prof. Dr. Denilson Alves Pereira.

Coorientador: Prof. Dr. André Pimenta Freire .

Bibliografia.

1. TCC. 2. Monografia. 3. Dissertação. 4. Tese. 5. Trabalho Científico – Normas. I. Universidade Federal de Lavras. II. Título.

CDD 808.066

ANNA PAULA FIGUEIREDO GONÇALVES

AVALIAÇÃO DE MODELOS DE LINGUAGEM PARA CRIAÇÃO DE *CHATBOTS* PARA A COMUNICAÇÃO COM PESSOAS SURDAS QUE USAM O PORTUGUÊS COMO SEGUNDA LÍNGUA

EVALUATION OF LANGUAGE MODELS FOR CREATING CHATBOTS FOR COMMUNICATION WITH DEAF PEOPLE WHO USE PORTUGUESE AS A SECOND LANGUAGE

Dissertação de Mestrado apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração Inteligência Artificial e Otimização, para a obtenção do título de Mestra.

APROVADA em 24 de Setembro de 2025.
Prof. Dr. Thiago Alexandre Salgueiro Pardo USP
Prof. Dr. Luiz Henrique de Campos Merschmann UFLA

gov.br
Documento assinado digitalmente
DENILSON ALVES PEREIRA
Data: 15/10/2025 17:04:47-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Denilson Alves Pereira
Orientador

Prof. Dr. André Pimenta Freire
Co-orientador

LAVRAS – MG
2025

Dedico este trabalho aos meus professores, pela sabedoria compartilhada; à comunidade surda, pela força e inspiração que motivaram cada etapa desta pesquisa; e à I.A, como símbolo de um futuro em que a tecnologia pode ampliar vozes e tornar o mundo mais acessível.

AGRADECIMENTOS

Agradeço a Deus pelo caminho trilhado até aqui e por todas as oportunidades concedidas. À minha família, pelo carinho. Às minhas amigas de longas datas Rafaela e Yasmine, pelo apoio mesmo que distante, às minhas amigas de casa Micaela, Gabrielle, Ingrid e Lorena , que tornaram meus dias mais alegres e continuam sendo parte importante da minha vida. Ao Mateus, pelo apoio emocional e por sempre acreditar no meu potencial e por sempre estar ao meu lado.

Agradeço à Érica Alves Barbosa e a Wanderson Samuel Moraes de Souza pelo suporte essencial com Libras e GLOSA, bem como pela introdução ao contexto da comunidade surda, sempre dispostos a compartilhar seus conhecimentos e experiências. Ao coorientador André Pimenta, pelo auxílio incansável na superação das questões burocráticas necessárias para a viabilização deste estudo.

Às secretárias, ao colegiado e aos colegas do LabRI, pelo suporte técnico, paciência e colaboração nas atividades cotidianas, pelo compartilhamento de ideias e pelo apoio nos momentos mais críticos do desenvolvimento deste trabalho.

Agradeço, ainda, aos professores com quem pude trocar experiências acadêmicas; ao Ahmed, pelo acompanhamento durante o estágio obrigatório; e, em especial, aos professores Luiz Henrique e Denilson, cuja exigência e orientação, embora desafiadoras, contribuíram de forma significativa para meu crescimento pessoal e profissional.

Por fim, expresso minha gratidão à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro, que foi fundamental para a realização desta pesquisa.

Educação não transforma o mundo. Educação muda pessoas. Pessoas transformam o mundo.

(Paulo Freire)

RESUMO

Modelos de linguagem de grande porte (LLMs) têm se consolidado como ferramentas poderosas para processamento de linguagem natural, mas seu desempenho ainda apresenta desafios quando aplicados a variantes linguísticas pouco representadas. A GLOSA — representação textual que reflete a estrutura sintática da LIBRAS — oferece um meio controlado para estudar como modelos lidam com construções linguísticas próximas da comunicação escrita de pessoas surdas, que frequentemente utilizam o Português como segunda língua. Este trabalho investigou a capacidade de adaptação de LLMs à GLOSA, com o objetivo de subsidiar o desenvolvimento de chatbots mais acessíveis à comunidade surda. Foram conduzidos experimentos comparativos envolvendo múltiplas arquiteturas abertas, de diferentes portes, avaliadas em cenários multilíngues (Inglês, Português, GLOSA e dados mistos). Duas estratégias de fine-tuning foram aplicadas a modelos pré-treinados, utilizando bases de dados construídos para refletir estruturas sintáticas reais da GLOSA. Os resultados confirmam perdas de desempenho ao migrar do Inglês para o Português e, principalmente, para a GLOSA, padrão já observado em estudos multilíngues. Contudo, o fine-tuning direcionado aumentou a interpretabilidade dos modelos, especialmente em arquiteturas maiores e com pré-treinamento multilíngue robusto, como Phi-4-14B e Qwen2.5-14B. Observou-se, porém, que ajustes excessivos ou desalinhados podem causar sobreajuste ou queda de acurácia, evidenciando a necessidade de controle rigoroso no processo de adaptação. Este estudo oferece três contribuições principais: (i) um conjunto inédito de dados em GLOSA e Português; (ii) um pipeline experimental reprodutível para avaliação de LLMs em cenários linguísticos não convencionais; e (iii) recomendações práticas para o ajuste de modelos em aplicações voltadas à acessibilidade.

Palavras-chave: LLMs; GLOSA; LIBRAS; chatbots; acessibilidade.

ABSTRACT

Large language models (LLMs) have established themselves as powerful tools for natural language processing, but their performance still presents challenges when applied to underrepresented linguistic variants. GLOSA—a textual representation that reflects the syntactic structure of LIBRAS—offers a controlled means of studying how models handle linguistic constructions similar to the written communication of deaf people, who frequently use Portuguese as a second language. This work investigated the adaptability of LLMs to GLOSA, with the aim of supporting the development of chatbots more accessible to the deaf community. Comparative experiments were conducted involving multiple open architectures of different sizes, evaluated in multilingual scenarios (English, Portuguese, GLOSA, and mixed data). Two fine-tuning strategies were applied to pre-trained models using databases constructed to reflect real GLOSA syntactic structures. The results confirm performance losses when migrating from English to Portuguese and, especially, to GLOSA, a pattern already observed in multilingual studies. However, targeted fine-tuning increased model interpretability, especially on larger architectures with robust multilingual pre-training, such as Phi-4-14B and Qwen2.5-14B. However, it was observed that excessive or misaligned adjustments can cause overfitting or decreased accuracy, highlighting the need for rigorous control in the adaptation process. This study offers three main contributions: (i) a novel dataset in GLOSA and Portuguese; (ii) a reproducible experimental pipeline for evaluating LLMs in unconventional linguistic scenarios; and (iii) practical recommendations for model tuning in accessibility-focused applications.

Keywords: LLMs; GLOSA; LIBRAS; chatbots; acessibility.

INDICADORES DE IMPACTO

A pesquisa contribui para a promoção da inclusão comunicacional da comunidade surda, investigando soluções computacionais capazes de compreender construções textuais que se aproximam da forma como surdos escrevem em português, frequentemente influenciadas pela estrutura da Libras. Embora o trabalho seja de natureza experimental, o desenvolvimento de modelos mais adequados ao GLOSA poderia potencialmente beneficiar milhares de estudantes e profissionais surdos que enfrentam barreiras de comunicação em ambientes educacionais, institucionais e de serviços públicos. A pesquisa se alinha aos ODS 4 (Educação de Qualidade) e 10 (Redução das Desigualdades) da ONU, ao propor tecnologias voltadas à acessibilidade linguística.

Conjuntos de dados bilíngues inéditos (GLOSA-Português) e um pipeline de experimentação reprodutível foram construídos, avaliando múltiplas arquiteturas de LLM (Phi-4-14B, Qwen2.5-14B, Llama-3.1-8B) sob diferentes estratégias de ajuste fino. Os resultados fornecem subsídios para o desenvolvimento de sistemas conversacionais com maior capacidade de interpretar variantes linguísticas controladas. Esse avanço poderá ser incorporado futuramente em produtos tecnológicos voltados à educação bilíngue, apoio ao ensino de Libras e serviços de chatbot acessíveis a pessoas com deficiência. O trabalho pode ser classificado principalmente nas áreas temáticas de Tecnologia e Produção (9), Educação (4) e Redução da Desigualdade (10).

Embora o estudo seja de natureza acadêmica e não envolva aplicação direta em comunidades externas durante esta fase, ele estabelece uma base sólida para futuras iniciativas de extensão. A inclusão de pessoas surdas na preparação e validação de dados — prevista como trabalho futuro — poderá impactar diretamente na formação de docentes, discentes e técnicos envolvidos, ampliando o alcance social do projeto. Em um cenário prático, os beneficiários diretos seriam escolas bilíngues, centros de apoio a pessoas com deficiência e desenvolvedores de tecnologias assistivas.

A pesquisa poderá fomentar parcerias institucionais entre universidades, centros de pesquisa e empresas do setor de tecnologia assistiva, fomentando a inovação com baixo custo adicional, visto que o uso de modelos abertos e ajustes controlados reduz barreiras financeiras. Além disso, reforça o papel da UFLA como promotora de soluções tecnológicas focadas em inclusão e acessibilidade, alinhadas às diretrizes nacionais dos programas de pós-graduação.

IMPACT INDICATORS

The research contributes to promoting communication inclusion for the deaf community by investigating computational solutions capable of understanding textual constructions that approximate the way deaf people write in Portuguese, often influenced by the structure of Libras. Although the work is experimental in nature, the development of models more suited to GLOSA could potentially benefit thousands of deaf students and professionals who face communication barriers in educational, institutional, and public service settings. The research aligns with the UN's SDG 4 (Quality Education) and SDG 10 (Reduced Inequalities) by proposing technologies aimed at linguistic accessibility.

Unprecedented bilingual datasets (GLOSA–Portuguese) and a reproducible experimentation pipeline were constructed, evaluating multiple LLM architectures (Phi-4-14B, Qwen2.5-14B, Llama-3.1-8B) under different fine-tuning strategies. The results provide support for the development of conversational systems with greater capacity to interpret controlled linguistic variants. This advancement can be incorporated in the future into technological products aimed at bilingual education, support for teaching Libras, and chatbot services accessible to deaf people. The work can be classified primarily under the thematic areas of Technology and Production (9), Education (4) and Inequality Reduction (10).

Although the study is academic in nature and does not directly involve application in external communities during this phase, it lays a solid foundation for future extension initiatives. The inclusion of deaf people in data preparation and validation—planned as future work—could have a direct impact on the training of faculty, students, and technicians involved, expanding the project's social reach. In a practical scenario, the direct beneficiaries would be bilingual schools, support centers for deaf people, and assistive technology developers.

The research could foster institutional partnerships between universities, research centers, and companies in the assistive technology sector, fostering innovation at low additional costs, as the use of open models and controlled adjustments reduces financial barriers. Furthermore, it reinforces UFLA's role as a promoter of technological solutions focused on inclusion and accessibility, aligned with national graduate program guidelines.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

ASL	*American Singing Language*
ENEM	Exame Nacional do Ensino Médio
ETILS	Estudos da Tradução e Interpretação da Língua de Sinais
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
INES	Instituto Nacional de Educação de Surdos
LBI	Lei Brasileira de Inclusão da Pessoa com Deficiência
Libras	Língua Brasileira de Sinais
LLaMA	Large Language Model Meta AI
LLM	Grande Modelo de Linguagem
LLMs	Grandes Modelos de Linguagem
LO	Linguagem Oral
LS	Linguagem de Sinal
LSTM	*Long Short-Term Memory*
MMLU	*Massive Multitask Language Understanding*
OMS	Organização Mundial da Saúde
PLN	Processamento de Linguagem Natural
PNS	Pesquisa Nacional de Saúde
RNA	Rede Neural Artificial
RNN	Redes Neurais Recorrentes
TIC	Tecnologia da Informação e Comunicação

SUMÁRIO

1 INTRODUÇÃO

A acessibilidade, delineada pela LBI (Lei Brasileira de Inclusão da Pessoa com Deficiência) (Lei nº 13.146/2015), constitui um conceito fundamental que estabelece princípios e diretrizes para a plena inclusão das pessoas com deficiência. Essa lei tem como finalidade garantir, em circunstâncias de equidade, liberdade e direitos, a inclusão social e a plena cidadania dessas pessoas. Essa legislação, promulgada em 2015, representa um marco normativo essencial no contexto brasileiro, consolidando avanços significativos como inclusão e acessibilidade. O embasamento legal proporcionado pela LBI delineia um arcabouço sólido para a implementação de políticas e práticas que visam a eliminação de barreiras e a promoção de ambientes acessíveis a todos os cidadãos (República, 2015).

A e-acessibilidade, uma abordagem apresentada pelo Governo Federal, concentra-se na eliminação de barreiras, destacando, em particular, o ambiente *online* (Digital, 2024; World Wide Web Consortium (W3C), 2023; International Organization for Standardization (ISO), 2016). Essa proposta procura assegurar que todos os cidadãos, independentemente de suas capacidades físico-motoras, perceptivas, culturais e sociais, possuam compreensão e controle efetivo na navegação de conteúdos e serviços governamentais.

A abordagem da e-acessibilidade ganha relevância significativa em um contexto onde as interações digitais desempenham um papel central na prestação de serviços públicos e na disseminação de informações. Nesse sentido, a promoção da acessibilidade na *web* não apenas se alinha com os princípios da inclusão, mas também reflete a evolução tecnológica e as demandas contemporâneas de uma sociedade cada vez mais digitalizada.

Uma projeção divulgada pela OMS (Organização Mundial da Saúde) estima que mais de um quarto da população mundial, o que corresponde cerca de 2,5 bilhões de pessoas, terá algum grau de perda auditiva até o ano de 2050 (Granda, 2021). A surdez, conforme definido pela própria OMS, refere-se à perda completa da capacidade de ouvir em um ou em ambos os ouvidos. Essa projeção tem implicações significativas na necessidade de disponibilização de tecnologias assistivas, uma vez que, de acordo com dados fornecidos pelas Nações Unidas, sintetizados por Sardenberg e Buogo (2022), quase 1 bilhão de pessoas com deficiência auditiva não têm acesso a nenhum tipo de tecnologia assistiva.

As tecnologias assistivas englobam uma variedade de recursos, dispositivos, equipamentos e sistemas desenvolvidos para promover maior independência e aprimorar a qualidade

de vida das pessoas com deficiência. Exemplos conhecidos dessas tecnologias incluem alguns equipamentos como cadeiras de rodas, próteses, rampas de acesso, sensores de movimento, relógios inteligentes com monitoramento de saúde, aplicativos e *softwares*. Essas tecnologias, quando acessíveis a pessoas com deficiência, facilitam na locomoção, comunicação e cognição (Sardenberg; Buogo, 2022). O propósito fundamental da aplicação dessas tecnologias é eliminar ou reduzir as barreiras que limitam a participação de indivíduos com deficiência em diversas atividades, tais como educação, trabalho, lazer e interação social.

A comunidade surda enfrenta desafios significativos, principalmente no âmbito da comunicação em Português, uma vez que o idioma é a segunda língua para muitos surdos (Araujo; Peixoto; Sousa, 2021). A principal barreira comunicacional entre surdos e ouvintes reside no fato de que a comunicação da comunidade surda se dá predominantemente através da Libras (Língua Brasileira de Sinais), reconhecida como meio legal de comunicação e expressão no Brasil desde a promulgação da LBI em 2015 (República, 2015).

Embora a Libras seja a língua oficial da comunidade surda brasileira, é importante destacar que, conforme estabelecido pela Constituição de 1988, a única língua oficial no Brasil é o Português (Constituinte, 1988). Essa dualidade linguística pode resultar em desafios adicionais para os alunos surdos, muitos dos quais não são fluentes em Português (Araujo; Peixoto; Sousa, 2021).

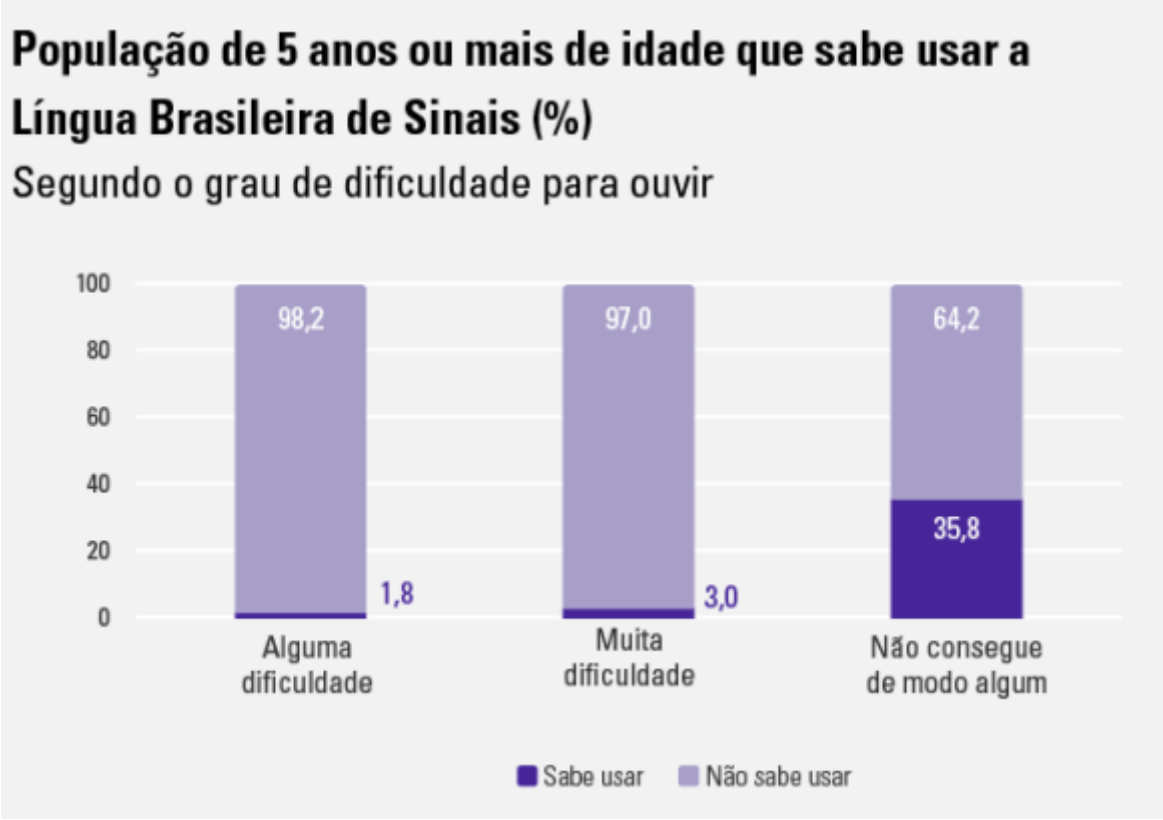
Os desafios enfrentados pela comunidade surda em relação à comunicação em Português são particularmente evidentes em situações práticas, como durante a realização do ENEM (Exame Nacional do Ensino Médio), onde a redação é elaborada em Português. Nesse contexto, a falta de fluência em Português torna-se um obstáculo significativo, afetando diretamente a habilidade dos alunos surdos em compreender e expressar-se por escrito durante a prova (Gandra, 2022).

Considerando os avanços tecnológicos na área do PLN (Processamento de Linguagem Natural), no que tange a interação humano-computador é predominantemente mediada pela linguagem humana, seja oral ou escrita, torna-se imperativo ponderar sobre a questão da acessibilidade durante o desenvolvimento de inovações nessa área. Por exemplo, ferramentas como o Google Tradutor, que realizam a tradução entre idiomas, e os *chatbots*, utilizados no atendimento ao cliente, ilustram a aplicação prática do PLN (Processamento de Linguagem Natural) na automatização de interações linguísticas. Dado o papel fundamental que esse campo desempenha na resolução de desafios cotidianos, tais como a tradução entre idiomas e a implemen-

tação de sistemas de respostas automáticas em diversos contextos, é importante assegurar que tais soluções sejam acessíveis e de fácil usabilidade para todos os usuários.

Segundo as estatísticas da PNS (Pesquisa Nacional de Saúde), apresentadas pela Estatísticas Sociais (2021) do IBGE (Instituto Brasileiro de Geografia e Estatística), cerca de 1,1% da população brasileira, o que corresponde a 2,3 milhões de indivíduos, enfrentam algum grau de deficiência auditiva. O estudo abrangeu pessoas com 2 anos ou mais, e revelou que, entre os brasileiros com idades entre 5 e 40 anos que apresentavam deficiência auditiva, apenas 22,4% eram familiarizados com a Libras. O estudo desdobrou essa familiaridade em níveis de audição, conforme a Figura 1.1, onde em análise detalhada, destaca ainda mais a dificuldade da comunidade surda em utilizar a Libras. Notavelmente, para os surdos que não escutam de forma alguma, nos quais deveriam se comunicar exclusivamente pela linguagem de sinais, cerca de 64,2% desses indivíduos não têm conhecimento da própria língua.

Figura 1.1 – Familiaridade em Libras por Nível de Audição



Fonte: Agência IBGE (2021).

O estudo identificou a presença de 31 mil crianças, com idades entre 2 e 9 anos, que enfrentam deficiência auditiva (Estatísticas Sociais, 2021). Essas crianças pertencem às gerações *Z* e *Alpha*, conhecidas por sua sólida conexão com a *internet* (Passero; Engster; Dazzi, 2016).

Essas gerações estão completamente envolvidas no uso cotidiano de dispositivos como *tablets*, *smartphones* e outras tecnologias, já que esses aparelhos são os principais meios de comunicação da atualidade. A comunicação por meio destes é quase que instantânea e permite o envio de áudios, vídeos e trocas de mensagens de texto. As mídias sociais, como *WhatsApp, Instagram e TikTok*, tornaram-se plataformas essenciais para a interação dessas gerações, nas quais oferecem meios diversificados para liberdade de expressão e compartilhamento de informações.

Além disso, a interação conversacional entre o ser humano e as máquinas é potencializada por tecnologias de PLN, que estreita a relação homem-máquina. São exemplos dessas tecnologias, o ChatGPT da OpenAI, o BERT do Google e o ELMO do Allen Institute for AI[1]

Segundo Taleb *et al.* (2021), a evolução das tecnologias tem possibilitado a automação de diversos processos, com perspectiva de expansão contínua. Os meios de comunicação se tornam mais eficientes com a implementação de tecnologias voltadas principalmente para simular o contato humano. Essas soluções, presentes em canais de autoatendimento, por exemplo, possibilitam conversas práticas e objetivas, reduz os custos operacionais e com pessoal. Além disso, é possível interagir de forma precisa e abrangente, de tal modo a oferecer uma conversa personalizada para cada perfil de cliente, ao utilizar de um linguajar familiar e individualizado, personalizado. Isso não apenas ajuda no processo de disseminar informações simples, precisas e de qualidade, mas também contribui para a eficácia do processo comunicativo.

Diversos *chatbots* estão disponíveis no mercado brasileiro, atendendo a múltiplas finalidades, desde atendimento ao cliente até funções de *marketing* ou bate-papos conversacionais. Alguns exemplos significativos incluem a LU, da Magazine Luiza; a AURA, da Vivo; a BIA, do Bradesco; e o *chatbot* de atendimento do Banco do Brasil. Cada um desses *chatbots* foi desenvolvido para atender às necessidades dos seus clientes e realizar diversas tarefas em suas respectivas áreas. Por exemplo, por meio da AURA, é possível efetuar pagamentos e consultar faturas, ao interagir com a LU, é possível rastrear encomendas, gerar boletos, entre outras funcionalidades. Já os *chats* do Banco do Brasil e a BIA são voltados para o contexto bancário, onde é possível fazer a verificação de saldos, realização de pagamentos e até mesmo transferências via PIX.

Entretanto, é importante ressaltar que essas soluções geralmente não são projetadas considerando o quesito acessibilidade (Araujo; Peixoto; Sousa, 2021). Conforme abordado pelo

[1] Mais detalhes sobre essas tecnologias podem ser encontrados nos sites oficiais: ChatGPT (https://www.openai.com/gpt), BERT (https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html), e ELMO (https://allennlp.org/elmo).

autor Haddon e Paul (2001), as soluções tecnológicas devem ser construídas para serem acessíveis, desde os primórdios de sua concepção. No escopo deste trabalho, voltado para indivíduos com alfabetização básica em Português, cuja língua primária é a Libras, o Português emerge como uma segunda língua, assim sendo, a acessibilidade um tópico a ser analisado e discutido. Os temas discutidos neste estudo destacam desafios consideráveis para o público que tem o Português como segunda língua, em especial o público que possui a Libras como primeira língua, que inclui um grande número de pessoas com baixo domínio do Português. O desafio é avaliar como essas pessoas interagem com recursos tecnológicos avançados na interação humano-computador, especificamente com os *chatbots*.

A Libras, enquanto uma LS (Linguagem de Sinal), apresenta complexidades eminentes, o que frequentemente demanda a intervenção de intérpretes para a tradução para o Português escrito e oral. Em contrapartida, os *chatbots* enfrentam desafios que incluem a necessidade de uma quantidade significativa de dados para seu desenvolvimento. Esses sistemas são caracterizados por sua alta complexidade, pois requerem recursos computacionais robustos para o desenvolvimento, além de uma base de dados com as características moduladas a um respectivo contexto, o que demanda extensos históricos de textos.

Aprofundar na compreensão desses desafios é importante para avançar no desenvolvimento de soluções mais eficazes e acessíveis, o que busca beneficiar tanto a comunidade surda quanto os usuários de *chatbots*, além das empresas e sociedade em geral. Essa análise aperfeiçoada não apenas revela as complexidades inerentes a essas áreas, mas também fornece *insights* valiosos que podem orientar o desenvolvimento de abordagens mais inovadoras e adaptáveis. A busca por soluções que interligam esses aspectos é fundamental para promover a inclusão e melhorar a experiência de usuários em ambas as esferas, contribuindo assim para melhorias significativas na interseção entre as linguagens e a interação humano-computador.

A GLOSA é uma forma de transcrição escrita que busca representar visualmente sinais da Libras utilizando elementos do Português escrito, sendo considerada uma ferramenta intermediária para conectar as duas línguas em contextos educacionais e de pesquisa. Conforme McCleary e Viotti (2007), seu principal objetivo é nivelar, em um formato comum, elementos da LIBRAS e do Português, de modo a facilitar o acesso e a compreensão entre usuários de diferentes sistemas linguísticos. A técnica consiste na grafia de palavras em letras maiúsculas, em que cada palavra corresponde a um sinal específico da LIBRAS, com significado aproximado ao termo em Português (Slobin, 2015). Um exemplo ilustra essa correspondência:

Português: Eu quero que o mundo tenha paz.

GLOSA: EU QUERER MUNDO P-A-Z[2].

Nesse caso, a estrutura sintática é adaptada à ordem e à lógica visual da LIBRAS, desconsiderando marcas flexionais e elementos funcionais do Português. No entanto, a GLOSA não reproduz a gramática da LIBRAS nem do Português, operando como uma variante escrita de função simbólica, e não linguística propriamente dita. Apesar de sua utilidade, autores como Leite et al. (2022) e Quadros, Soares e Miranda (2014) destacam limitações importantes, como a ausência de padronização, a dificuldade de representar aspectos visuais, emocionais e contextuais da LIBRAS, e a escassez de sistemas tecnológicos adaptados para a escrita sinalizada. Mesmo com o avanço de mídias digitais e ferramentas multimodais, a prática da GLOSA persiste como estratégia predominante, evidenciando tanto sua relevância quanto os desafios relacionados à representação escrita das línguas de sinais.

1.1 Objetivos

Este trabalho tem como objetivo geral avaliar a viabilidade de utilizar modelos de linguagem de grande porte para interpretar textos em GLOSA, tendo em vista que a escrita em Português produzida por muitas pessoas surdas frequentemente apresenta estruturas sintáticas semelhantes à GLOSA, por refletirem diretamente a organização dos sinais da LIBRAS. Essa aproximação entre a GLOSA e o Português escrito pelos surdos, apontada por estudos como (McCleary; Viotti, 2007; Nascimento; Costa, 2022), justifica o foco na GLOSA como representação intermediária para subsidiar o desenvolvimento de *chatbots* mais acessíveis e eficazes para a comunidade surda, que utiliza o Português como segunda língua. Os objetivos específicos são:

1. Avaliar soluções de *chatbots* disponíveis, focando na identificação da familiaridade dessas ferramentas com o modo de escrita dos usuários da comunidade surda.

2. Verificar a capacidade de modelos de linguagem na interpretação de textos na GLOSA, sem *fine-tuning*, identificando as principais limitações.

[2] Referência palavra “PAZ” consultado no Dicionário de LIBRAS do Instituto Nacional de Educação de Surdos (INES), disponível em: https://www.ines.gov.br/dicionario-de-LIBRAS/

3. Investigar o impacto do *fine-tuning* em diferentes configurações na adaptação dos modelos para compreender textos em GLOSA.

4. Comparar os resultados obtidos entre diferentes modelos e estratégias de treinamento, propondo recomendações para o uso em *chatbots* voltados a pessoas surdas.

1.2 Justificativa

Este estudo parte da necessidade de compreender e superar barreiras específicas de acessibilidade digital enfrentadas pela comunidade surda, especialmente no uso de *chatbots* e interação com LLMs (Grandes Modelos de Linguagem). Com o avanço acelerado das TIC (Tecnologia da Informação e Comunicação) e o uso crescente de modelos de linguagem, surge uma oportunidade real de oferecer soluções mais inclusivas, capazes de respeitar as particularidades linguísticas dessa população.

A literatura aponta que o Português escrito na forma GLOSA — uma variante controlada e simplificada da Libras — é fundamental para viabilizar a comunicação entre pessoas ouvintes e não ouvintes (Leite *et al.*, 2022; Nascimento, 2023). No entanto, essa área ainda é incipiente, e enfrenta desafios como a padronização da escrita e a ausência de soluções que interpretem adequadamente suas estruturas sintáticas e semânticas. Trabalhos relacionados ao contexto da TV 3.0 (Costa *et al.*, 2023) já indicam caminhos promissores para adaptar conteúdos em Português para GLOSA, mas também reforçam a necessidade de ampliar essa abordagem para outras tecnologias interativas, como os *chatbots*.

O trabalho de revisão da literatura apresentado por Khennouche *et al.* (2023) destaca o crescimento exponencial de pesquisas com *chatbots* na última década, mas também evidenciam lacunas importantes quanto à sua adaptação para públicos que utilizam o Português como segunda língua, como por exemplo, a falta de referências gramaticais. Além do trabalho apresentado por Goldman *et al.* (2025) que aborda uma análise qualitativa da interação de pessoas Surdas com os LLMs no contexto inglês. Conforme discutido no Capítulo 4, poucos estudos avançaram para além da tradução automática básica entre idiomas, e praticamente não há registros de modelos de linguagem treinados especificamente para reconhecer e gerar textos em GLOSA.

Diante disso, este trabalho propõe-se a investigar a viabilidade de utilizar LLMs ajustados por *fine-tuning* para interpretar essa variante linguística, abordagem que ainda não foi

identificada na literatura revisada. Além de avaliar o desempenho dos modelos em diferentes cenários (com e sem ajuste fino e com distintas combinações de conjuntos de dados), busca-se fornecer subsídios concretos para o desenvolvimento de chatbots que atendam de forma mais precisa e eficaz à comunidade surda.

Essa proposta está em consonância com os princípios da LBI e amplia o debate sobre tecnologias realmente acessíveis. Ao tornar a interação digital mais compreensível para pessoas que usam GLOSA, contribui-se não apenas para o acesso à informação, mas também para a construção de uma sociedade mais justa e inclusiva, fortalecendo oportunidades educacionais, profissionais e comunicacionais para esse público.

1.3 Organização do trabalho

O trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta o referencial teórico, abordando os principais conceitos relacionados ao projeto. Em seguida, no Capítulo 3, são descritas as metodologias empregadas para a execução do projeto proposto. O Capítulo 4 detalha a estrutura lógica adotada para a coleta de trabalhos correlatos, além de apresentar os principais estudos encontrados. O Capítulo 5 apresenta uma análise prévia de *chatboots* disponíveis no mercado, para identificação de quão bom eles são para interpretar a GLOSA. No Capítulo 6, são expostos o desenvolvimento do trabalho, as etapas realizadas, os recursos computacionais e tecnologias utilizadas, bem como a modelagem e os métodos de avaliação. Posteriormente o capítulo é sucedido pelo Capítulo 7, que apresenta e discute os resultados obtidos. Por fim, o Capítulo 8 traz uma síntese dos resultados, destacando as contribuições do trabalho e apontando perspectivas para pesquisas futuras.

2 REFERENCIAL TEÓRICO

Cada estudo é conduzido em um contexto único, com suas próprias diferenças e terminologias específicas. Portanto, para contextualizar sobre o universo de estudo deste trabalho, o referencial teórico permite explorar os principais termos que permeiam essa pesquisa. Os termos aqui apresentados são mais que simples palavras, eles representam conceitos fundamentais que moldam e direcionam a investigação sobre a acessibilidade e tecnologia. O estudo aborda detalhes sobre uma gama de assuntos, nos quais são inclusos o Português como segunda língua, a comunidade surda, os intérpretes, GLOSAS, Libras, e conceitos relacionados ao PLN, como, por exemplo, assistentes virtuais inteligentes, redes neurais, modelos de linguagem, *tokens* e LLMs.

Cada um desses elementos compõe um quadro teórico abrangente que contribui para uma compreensão aprofundada dos desafios e oportunidades envolvidos na acessibilidade tecnológica para indivíduos que possuem o Português como L2. Uma compreensão abrangente de cada conceito capacita a desenvolver uma análise robusta e amplamente fundamentada. A estratificação do referencial teórico não apenas amplia o entendimento da temática, mas também inspira a explorar novos caminhos e contribuir para o avanço do conhecimento.

2.1 Português como segunda língua

O Português desempenha o papel de segunda língua para uma variedade de pessoas, abrange desde pessoas com deficiências ou problemas cognitivos até aquelas que estão aprendendo a língua como estrangeira. Contudo, o enfoque deste trabalho incide sobre um grupo específico: aquele que, por ter a Libras como linguagem primária, utiliza o Português como segunda língua, ou L2.

A população surda é geralmente reconhecida como bilíngue, tendo a Libras como sua primeira língua e o Português como segunda língua (L2), predominantemente na modalidade escrita (Araujo; Peixoto; Sousa, 2021). Essa dinâmica bilíngue é marcada por assimetrias no acesso e na circulação das línguas. Como apontam Leite *et al.* (2022), fatores sociolinguísticos, como a menor visibilidade e institucionalização da Libras, afetam diretamente o desenvolvimento do letramento em Português por parte dos surdos. A LIBRAS, por exemplo, ainda não possui o mesmo grau de circulação social e institucional que o Português, sendo pouco utilizada

em ambientes escolares, jurídicos e acadêmicos (Quadros; Soares; Miranda, 2014). Soma-se a isso a ausência de um sistema de escrita amplamente adotado para a língua de sinais, o que representa uma barreira significativa no acesso à informação e à comunicação mediada por texto nas mídias contemporâneas (Araujo; Peixoto; Sousa, 2021).

Portanto, é essencial compreender os elementos que constituem essa dinâmica bilíngue, a fim de identificar e atender às necessidades específicas da população surda. Essa dinâmica envolve não apenas o uso de duas línguas distintas — LIBRAS e Português — mas também os contextos sociais, institucionais e culturais em que essas línguas circulam. Para contextualizar esses aspectos de forma mais ampla, apresentam-se a seguir algumas definições e conceitos que contribuem para a compreensão dos principais agentes envolvidos nos processos comunicativos com o público-alvo: as pessoas surdas, as pessoas com deficiência na fala (mudas), os ouvintes e os intérpretes de LIBRAS.

Entretanto, é pertinente a observação de Leite *et al.* (2022), que destaca a complexidade do bilinguismo na comunidade surda, adicionando novas camadas à dinâmica linguística discutida. Os autores argumentam que, embora seja importante valorizar e fortalecer o uso da LIBRAS, excluir o Português dessas discussões seria uma abordagem limitada. Isso porque o contato com a língua portuguesa, ainda que na modalidade escrita e muitas vezes com dificuldades, é constante e inevitável no cotidiano das pessoas surdas. Tal convivência linguística, marcada por assimetrias de poder e acesso, constitui parte integrante da construção identitária e cultural da comunidade surda, exigindo, portanto, abordagens educativas e tecnológicas que reconheçam essa complexidade.

2.2 Língua Brasileira de Sinais - LIBRAS

A Libras desempenha um papel fundamental como principal sistema de comunicação na comunidade surda, respaldada pela Lei nº 10.436, de 24 de abril de 2002, que a reconhece oficialmente, como:

> Entende-se como Língua Brasileira de Sinais a forma de comunicação e expressão, em que o sistema linguístico de natureza visual-motora, com estrutura gramatical própria, constitui um sistema linguístico de transmissão de ideias e fatos, oriundos de comunidades de pessoas surdas do Brasil.

A Libras é uma linguagem visual-motora que se baseia na comunicação por gestos e visualização. Esses gestos são articulados por meio das mãos, braços, expressões faciais, mo-

vimentos dos lábios e uso do espaço, sendo percebidos e interpretados pelos usuários da língua através da visão (Leite *et al.*, 2022).

Apesar do reconhecimento legal da Libras como o sistema linguístico específico da comunidade surda, a lei estipula em seu artigo final, que ela não pode substituir o Português escrito, perpetuando assim a hegemonia do Português sobre a Libras. A disparidade entre a Libras e o Português é evidenciada principalmente na prática recorrente de explicar os sinais por meio de palavras em Português, uma estratégia conhecida como "GLOSSAGEM" (Leite *et al.*, 2022; Nascimento, 2023). Contudo, é importante destacar que essa prática não possui um registro formalizado. Essa dualidade entre a Libras e o Português evidencia a necessidade contínua de conscientização e discussão sobre a igualdade linguística e cultural para a comunidade que utiliza a Libras como língua primária e o Português como secundária (Leite *et al.*, 2022).

2.2.1 Surdo

De acordo com o Decreto nº 5.626, que regulamenta a Lei nº 10.436, de 24 de abril de 2002, o estado de surdez é caracterizado da seguinte forma:

> Art. 2o Para os fins deste Decreto, considera-se pessoa surda aquela que, por ter perda auditiva, compreende e interage com o mundo através de experiências visuais, manifestando sua cultura principalmente pelo uso da Língua Brasileira de Sinais - LIBRAS.

Desse modo, é consolidado o principal meio de comunicação dos surdos como sendo através da LS. Assim como os idiomas, cada nacionalidade possui suas próprias características e significados em seu dialeto, e com a LS não é diferente, assim sendo totalmente restritas e únicas ao seu berço. Por exemplo, no Brasil têm-se a Libras, já nos Estados Unidos a LS é a *American Sign Language*, elas são línguas totalmente independentes e cada uma possui sua própria gramática, vocabulário e cultura associada.

2.2.2 Mudo

Uma pessoa muda é aquela que, no que lhe concerne pode não possuir a habilidade de produzir sons da fala, diferenciando-se da pessoa surda, que, embora possua a capacidade auditiva, opta por não utilizar o aparelho fonador para expressar-se vocalmente (Soleman; Bousquat, 2021). Infelizmente, é comum a confusão entre esses dois termos, mesmo sendo primordial

compreender que surdez e mudez são deficiências distintas. A grande maioria dos surdos possui as cordas vocais plenamente funcionais, sendo a mudez uma condição menos comum entre eles (Soleman; Bousquat, 2021).

As pessoas mudas, também utilizam-se da Libras como uma língua presente em seu meio de comunicação. Além desse modo de comunicação, é possível também a leitura labial, entretanto o que vem ganhando espaço é a escrita, já que ela é uma modalidade especialmente relevante na era digital, onde a comunicação por meio de texto é predominante.

2.2.3 Ouvinte

No contexto da comunicação por sinais, o termo "ouvinte" é utilizado para designar pessoas que têm a capacidade tanto de falar quanto de ouvir. Contudo, é importante destacar que pessoas ouvintes não estão restritas apenas à comunicação oral; elas também podem recorrer à LS para interagir com indivíduos que a utilizam como principal forma de expressão. Além disso, para aqueles que enfrentam limitações na fala, como as pessoas mudas, a comunicação por meio de sinais representa uma alternativa eficaz.

Assim, é importante reconhecer que a comunicação por sinais, não se restringe apenas à comunidade surda, mas também pode ser uma forma eficaz de expressão para pessoas ouvintes que enfrentam desafios na fala. Essa compreensão mais ampla contribui para uma visão inclusiva e abrangente da diversidade de modos de comunicação existentes.

2.2.4 Tradutores e Intérpretes

O papel fundamental dos intérpretes e tradutores de uma LS é estabelecer uma ponte entre as pessoas que têm uma língua de sinais — como a Libras — como primeira língua e os ouvintes, possibilitando o acesso à educação, a bens e serviços públicos e, sobretudo, à informação. Essa função é essencial para assegurar que essas pessoas tenham os mesmos direitos e oportunidades de comunicação que os demais cidadãos (Nascimento, 2016)

O intérprete trabalha com a linguagem falada ou de sinais, e reproduz em tempo real o que está sendo dito pelo emissor da mensagem aos demais envolvidos. Segundo Rodrigues (2018), o intérprete de Libras precisa ter raciocínio rápido e ser fluente, pois, precisa transmitir uma mensagem quase que de forma instantânea. A atuação do intérprete é dinâmica, pois

exige de movimentos e expressões faciais e corporais contextualizadas para garantir uma boa comunicação. Como na Libras utiliza-se dessas expressões físicas, elas podem sofrer influências externas e do próprio ambiente, assim como nas variações linguísticas, como por exemplo a regionalidade, pode haver influências por parte dos intérpretes, o que pode gerar uma entrega de mensagem equivocada (Rodrigues, 2018).

Já o tradutor, trabalha com a linguagem escrita, convertendo textos de um idioma para outro (Rodrigues, 2018). No entanto, dado o contínuo crescimento e desenvolvimento do campo da Libras, existem algumas técnicas nas quais é possível traduzir os sinais para a forma escrita. O registro dessa escrita é feito por meio do uso de imagens e também utilizando o alfabeto, através de textos em Português. Dois modos conhecidos dessas técnicas são o *singing writing* e GLOSAS. *Siging writing* é a técnica de escrever através de símbolos, já a GLOSA é através dos caracteres.

Conforme apresentado por Nascimento (2023) o ETILS (Estudos da Tradução e Interpretação da Língua de Sinais) têm ganhado destaque no Brasil, refletindo o crescente interesse e valorização da Libras. Todavia, o campo enfrenta desafios relacionados à complexidade da tradução e interpretação da LS, pois ela abrange desde habilidades motoras até competência linguística e comunicativa dos envolvidos (Nascimento, 2023). Essa necessidade é apresentada por Nascimento (2023), onde é sugerido, que o uso de TICs podem ser integradas e disponibilizadas para este público a fim de enfrentar os desafios conceituais, contextuais e de comunicação em benefício da comunidade em questão.

2.2.5 GLOSA

De acordo com McCleary e Viotti (2007), a GLOSA é uma ferramenta essencial para estabelecer uma conexão visual entre textos de duas línguas, em especial neste contexto entre a Libras e o Português. Segundo os autores, seu principal propósito é nivelar as duas linguagens em um formato comum, isto é, que seja capaz de representar ambas. A GLOSA consiste em palavras grafadas em maiúsculo, onde cada uma delas representa um sinal, esse sinal, possui significado equivalente em Português (Slobin, 2015).

No âmbito da pesquisa sobre LS, McCleary e Viotti (2007) apontam a tendência inevitável de associar semanticamente e gramaticalmente as palavras do Português aos sinais da Libras, o que pode resultar em diversas confusões, uma vez que a Libras sofre influências ex-

ternas do ambiente e do próprio emissor e do intérprete. Nesse contexto, as GLOSAS são ponderadas como nomes convencionais associados aos sinais, com o intuito de evitar equívocos quanto ao significado do sinal ou sua função gramatical. Além disso, as diferenças na prática de transcrição e "GLOSSAGEM", em comparação com a LO (Linguagem Oral), podem levar a problemas na representação e interpretação dos sinais e texto, uma vez que a transcrição pode ser afetada pela omissão de elementos essenciais na comunicação oral, como sentimentos, emoções e contexto ambiental (Leite *et al.*, 2022).

Um ponto crítico proeminente é a persistência na representação das LSs por meio de palavras escritas, mesmo em uma era cercada por avanços tecnológicos. Apesar da disponibilidade de recursos como fotos, vídeos e a própria escrita de sinais, a prática do uso de GLOSAS permanece, especialmente devido à falta de adaptação dos sistemas de escrita de sinais para o contexto de mídias digitais (Quadros; Soares; Miranda, 2014; Leite *et al.*, 2022). Os autores Quadros, Soares e Miranda (2014) chamam a atenção para a falta de padronização nesse aspecto e a escassez de evolução tecnológica, principalmente com os avanços das TICs, evidenciando a necessidade de abordagens mais inovadoras ao estudar e representar as LSs.

É importante ressaltar que, assim como as línguas de sinais variam de acordo com cada país, os modos de escrita em GLOSA também apresentam diferenças. A técnica empregada é a mesma — a GLOSSAGEM —, mas a forma escrita difere conforme a língua de sinais de referência. Na 2.1, apresentada por Othman e Jemni (2011), observa-se um diálogo em Língua de Sinais Americana (ASL), no qual fica evidente a diferença entre a escrita de uma frase em Inglês e sua correspondente em GLOSA. De maneira análoga, na Figura 2.2 é exibido um exemplo extraído do dicionário de dados do INES, que ilustra a escrita em GLOSA para a palavra BANCO. [1].

[1] Esse material pode ser consultado em: http://www.acessibilidadebrasil.org.br/libras_3/

Figura 2.1 – Exemplo Diálogo Entre Surdos - Inglês

A: (get-attention) TOMORROW I GO PICK-up BOOK NEW I BUY YOU DON'T-MIND I BORROW YOUR TRUCK?
"Tomorrow I'm going to pick up some new books I just bought. Do you mind if I borrow your truck?"
B: TOMORROW TIME WHAT?
"What time tomorrow?"
A: AROUND 10 "give-or-take"
"Maybe around 10."
B: NO NOT WORK MY TRUCK me-BRING MECHANIC FIX TOMORROW MORNING. TOMORROW AFTERNOON BETTER.
"No, that won't work; I need to take the truck to get serviced tomorrow morning. The afternoon would work better."
A: FINE. YOU 2 TOMORROW FINE ?
"That's fine. Would 2 work for you?"
B: SURE-SURE FINE.
"Yes, that works fine."

Fonte: Othman (2011).

Figura 2.2 – Exemplo Frase - Português

Exemplo	Exemplo Libras
Agora eu vou ao banco porque preciso pagar a conta de luz.	AGORA EU IR BANCO PRECISAR PAGAR LUZ.
Classe Gramatical	**Origem**
SUBSTANTIVO	nacional

Fonte: INES (2025).

2.3 Processamento de Linguagem Natural

O PLN é o campo da IA que centraliza a interação entre computadores e humanos, por meio da própria linguagem humana. O objetivo do PLN é permitir que os computadores compreendam, interpretem e gerem texto ou fala equitativamente como os seres humanos (Allen, 1995; Jurafsky; Martin, 2019). Segundo Liddy (2001), são exemplos que envolvem o PLN:

recuperação de informação a partir de textos, tradução automática, interpretação de textos e a realização de inferências a partir de uma base de dados textual.

O PLN por ser um campo multidisciplinar, engloba uma variedade de aspectos linguísticos tais como fonética e fonologia, morfologia, sintaxe, semântica e pragmática. Uma compreensão aprofundada desses elementos é fundamental para a análise e interpretação da linguagem em diversos contextos comunicativos.

2.3.1 Aspectos Linguísticos do PLN

Ao longo das últimas décadas, diversos pesquisadores têm investigado de forma aprofundada os aspectos linguísticos e comunicacionais relacionados ao bilinguismo e à acessibilidade de pessoas surdas. Trabalhos como os de Allen (1995), Manning e Schütze (1999) e Liddy (2001) oferecem contribuições relevantes para esse campo de estudo, cujas principais percepções e abordagens podem ser sintetizadas a seguir:

- Fonéticos e Fonológicos: envolve o processamento dos sons da linguagem falada e sua representação em forma escrita. Compreende a análise dos fonemas, isto é, unidades sonoras distintivas e sua correspondência com os grafemas, ou seja, as unidades de escrita. Desse modo é possível converter a fala em texto escrito, identificando os fonemas em uma palavra e sua respectiva ortografia.

- Morfológicos: a análise morfológica diz respeito à estrutura e formação das palavras, inclui a identificação de prefixos, sufixos, raízes e outras unidades significativas. Dessa forma, é possível identificar as diferentes formas de uma palavra com base em sua raiz e afixos, como, por exemplo: “cantar”, “cantava”, “cantaria”.

- Sintáticos: a análise sintática se concentra na estrutura gramatical das frases e na relação entre as palavras para determinar sua função na sentença. É útil ao analisar a estrutura de uma frase para identificar sujeito, verbo e objeto, além de determinar a ordem das palavras.

- Semânticos: na semântica, o foco é compreender o significado das palavras e frases, além de considerar o contexto e as relações entre elas. Essa análise é importante pois permite interpretar o significado de uma frase ou palavra com base em seu contexto, como, por

exemplo, a palavra “banco” pode ser interpretada como uma instituição financeira ou um assento.

- Pragmáticos: o estudo pragmático concentra-se no uso prático da linguagem em diferentes situações comunicativas, considerando o contexto, a intenção do falante e as inferências necessárias para compreender o significado por completo.

A análise desses aspectos visa entender não apenas as palavras isoladas, mas também a estrutura completa das interações verbais em uma variedade de ambientes, que inclui conversas cotidianas, textos escritos, mídias sociais e outras formas de comunicação verbal e textual.

2.3.2 Aplicações do PLN

Os estudos que envolvem o PLN é caracterizado por sua complexidade e constante mutabilidade, tendo experimentado avanços significativos, impulsionando uma gama de aplicações em diversos campos. Desde assistentes virtuais até sistemas de tradução automática, o PLN desempenha um papel fundamental na automatização e melhoria da interação humano-computador. A seguir, são abordadas algumas dessas aplicações para contextualizar a importância e a diversidade desse campo em constante evolução.

Os assistentes virtuais, como a *Siri* da *Apple*, a *Alexa* da *Amazon* e o *Google Assistant*, são exemplos proeminentes de aplicações práticas do PLN. Esses sistemas utilizam técnicas avançadas para entender e responder a comandos de voz dos usuários. Por exemplo, ao solicitar a previsão do tempo ou agendar compromissos, esses assistentes interpretam a fala do usuário, extraem informações relevantes e executam ações correspondentes (Moraes, 2023.).

A tradução automática de idiomas é uma aplicação amplamente conhecida, é comumente encontrada em plataformas como o *Google Translate* e o *DeepL* (Hendy *et al.*, 2023). Essas ferramentas empregam modelos de linguagem avançados para traduzir texto de uma língua para outra de forma rápida e precisa. Esses sistemas utilizam técnicas de PLN para analisar e compreender o significado das sentenças em um idioma fonte, para assim, produzir uma tradução equivalente no idioma alvo.

Com o crescimento explosivo das mídias sociais, a análise de sentimentos também se tornou uma aplicação relevante no campo do PLN. Empresas e pesquisadores utilizam algoritmos para analisar grandes volumes de postagens em mídias sociais e identificar tendências de

sentimentos em relação aos produtos, eventos ou algum tópico específico. Por exemplo, ferramentas de análise de sentimentos podem ser usadas para monitorar a satisfação do cliente com base em postagens em redes sociais sobre um produto ou serviço (AL-Ghuribi; Noah, 2021).

Essa análise permite com que as empresas compreendam melhor o *feedback* dos clientes, monitorar a reputação da marca e identificar oportunidades de melhoria, resultando em estratégias de negócios direcionadas. Além disso, ao identificar padrões de sentimentos dos consumidores, as empresas podem adaptar suas campanhas de *marketing* e desenvolver produtos ou serviços que atendam às necessidades e desejos do público-alvo. Essas informações também podem ser utilizadas para antecipar crises de reputação e responder de forma proativa, o que fortalece a relação com os clientes e constrói uma imagem positiva da marca no mercado.

2.4 Inteligência Artificial

A IA (Inteligência Artificial) tem desempenhado um papel central na transformação de processos comunicacionais, educacionais e tecnológicos nas últimas décadas. Seu desenvolvimento progressivo tem ampliado fronteiras em áreas como a compreensão de linguagem natural, a geração automatizada de conteúdos e a personalização de experiências de aprendizagem. Nesta seção, apresenta-se um panorama conceitual e técnico da IA, com ênfase em suas aplicações no campo da linguagem e em suas implicações para a acessibilidade e inclusão digital.

A abordagem inicia-se com uma introdução às Redes Neurais Artificiais, estruturas fundamentais para o funcionamento de sistemas inteligentes contemporâneos. Em seguida, explora-se a IA Generativa, categoria que abrange modelos capazes de criar textos, imagens ou outros dados a partir de padrões aprendidos. A subseção 2.4.3 dedicada aos Modelos de Linguagem detalha os componentes internos dessas arquiteturas, como *tokens* e *embeddings*, essenciais para a representação e manipulação da linguagem natural por máquinas.

Avançando nesse escopo, são discutidos os LLMs, com destaque para seus potenciais e limitações em contextos educativos e comunicacionais. Na sequência, apresenta-se o conceito de *Prompt Engineering*, técnica fundamental para orientar e refinar as respostas geradas por esses modelos. A subseção 2.4.6 também contempla uma discussão sobre técnicas de *fine-tuning*, que viabilizam a adaptação de LLMs a contextos e públicos específicos.

Essa organização visa oferecer uma base conceitual sólida para compreender os mecanismos subjacentes à IA moderna, bem como suas aplicações práticas em contextos de acessibilidade, com ênfase no público surdo e em práticas pedagógicas mediadas por tecnologia.

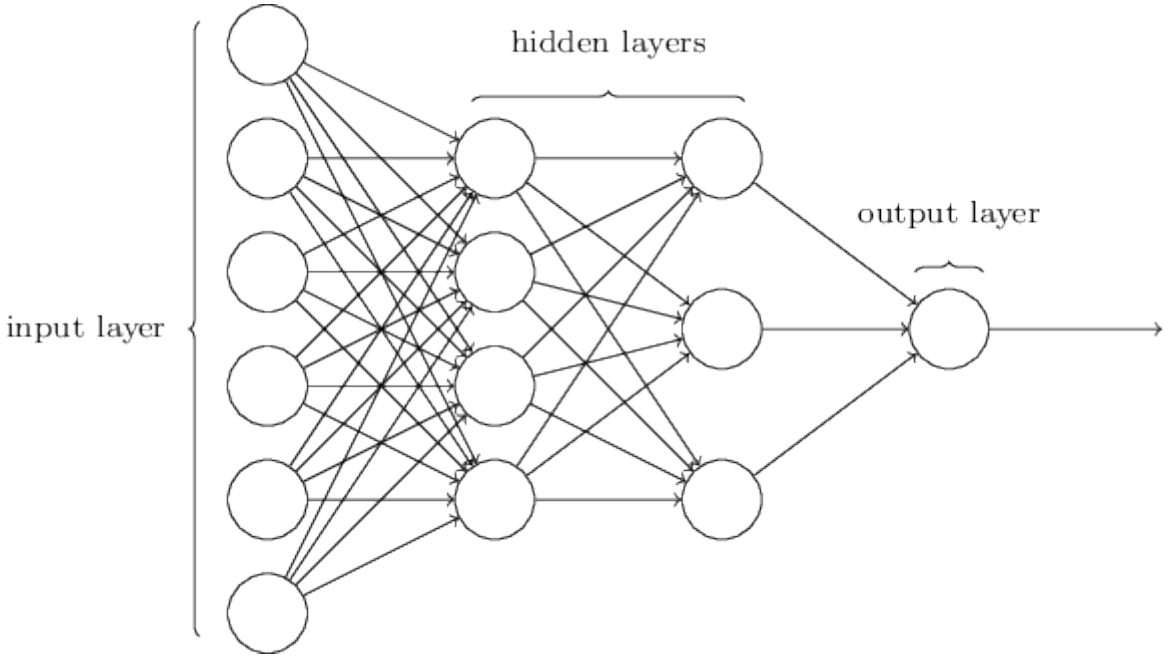
2.4.1 Redes Neurais Artificiais

Uma RNA (Rede Neural Artificial) é um modelo computacional inspirado no funcionamento de uma das partes do sistema nervoso central, o cérebro humano. Essas redes constituem um dos principais fundamentos da IA (Inteligência Artificial), desempenhando um papel central nos algoritmos de PLN. A concepção estrutural da RNA busca imitar a forma como os neurônios biológicos interagem para processar e transmitir informações (Allen, 1995; Braga; Ludermir; Carvalho, 2000).

A arquitetura de uma RNA é composta por múltiplas camadas de nós interconectados, que simulam a organização dos neurônios no cérebro biológico conforme apresentado na Figura 2.3. Essa estrutura é composta por três tipos principais de camadas:

- **Camada de entrada**: Responsável por receber os dados brutos e transmiti-los para as camadas subsequentes da rede.
- **Camadas ocultas**: Uma ou mais camadas intermediárias onde ocorre o processamento e a extração de padrões complexos a partir dos dados de entrada. Cada neurônio nessas camadas aplica funções de ativação e ajusta seus pesos conforme o aprendizado da rede.
- **Camada de saída**: Fornece a previsão final da rede neural, dependendo da tarefa específica, como classificação, regressão ou geração de texto.

Figura 2.3 – Arquitetura de uma Rede Neural Artificial



Fonte: Deep Learning Book Brasil (2024).
Disponível em: https://www.deeplearningbook.com.br/a-arquitetura-das-redes-neurais/

Os neurônios dentro dessas camadas computam valores de entrada e os transmitem por meio de conexões ponderadas. Cada conexão entre neurônios é associada a um peso, que define a importância do sinal transmitido. Durante o treinamento, esses pesos são ajustados por meio de algoritmos de otimização, como o *backpropagation*, o que permite que a rede aprimore seu desempenho em tarefas específicas.

As RNAs são amplamente aplicadas em diversos campos da ciência da computação e engenharia, principalmente em problemas que envolvem o reconhecimento de padrões complexos, como:

- **PLN**: Aplicações como tradução automática, geração de texto e análise de sentimentos se beneficiam da capacidade das redes neurais de modelar padrões linguísticos complexos.
- **Visão Computacional**: Redes neurais convolucionais (CNNs) são utilizadas para reconhecimento de imagens, detecção de objetos e segmentação semântica.
- **Sistemas de Recomendação**: Algoritmos baseados em redes neurais auxiliam na personalização de conteúdos em plataformas de *streaming* e comércio eletrônico.
- **Modelagem de Séries Temporais**: Redes neurais recorrentes (RNNs) e *transformers* são empregados para previsões em áreas como finanças, meteorologia e biomedicina.

Dessa forma, as RNAs desempenham um papel essencial no avanço da inteligência artificial moderna, permitindo que sistemas computacionais processem informações de maneira mais eficiente e realizem tarefas com precisão aprimorada.

2.4.2 IA Generativa

O termo Inteligência Artificial Generativa (IA Generativa) surgiu com a evolução de modelos capazes de produzir conteúdos inéditos a partir de grandes volumes de dados. A expressão começou a ganhar destaque principalmente após os avanços dos modelos de redes neurais generativas, como Redes Adversárias Generativas (GANs) (Goodfellow *et al.*, 2014), e posteriormente com a popularização dos *transformers* aplicados ao PLN, como o GPT (Generative Pre-trained Transformer) (Radford *et al.*, 2018a).

A IA Generativa se conecta diretamente ao PLN porque os modelos avançados de linguagem, como GPT, BERT, T5 e Gemini, não apenas analisam e interpretam texto, mas também são capazes de gerar conteúdos originais com base em padrões aprendidos em grandes conjuntos de dados textuais. Esses modelos utilizam arquiteturas de Transformers, que revolucionaram o PLN ao permitir a modelagem eficiente de dependências de longo alcance entre palavras e frases em um texto (Vaswani *et al.*, 2017).

2.4.3 Modelos de Linguagem

Os modelos de linguagem são sistemas computacionais treinados em grandes volumes de dados para identificar padrões e produzir saídas, como prever a próxima palavra em uma sequência de texto ou gerar imagens em problemas gráficos. No campo linguístico, ele é utilizado para prever a probabilidade de ocorrência de palavras ou sequências de palavras em um contexto textual. Esses modelos desempenham um papel central no PLN, viabilizando diversas aplicações (Jurafsky; Martin, 2019).

Um modelo de linguagem pode ser definido como um modelo probabilístico que representa estatisticamente a forma como palavras e outros elementos linguísticos se relacionam em um idioma (Jurafsky; Martin, 2019). Durante o treinamento esses modelos ajustam seus parâmetros internos para otimizar a previsão de palavras subsequentes, aprimorando sua capacidade de compreender e gerar texto de maneira coerente e contextualizada.

A construção de modelos de linguagem frequentemente envolve o uso de RNA, que são capazes de identificar padrões complexos em dados textuais (Andreas; Vlachos; Clark, 2013; Pham *et al.*, 2013). Entre essas abordagens, destaca-se o uso de arquiteturas baseadas em *Transformers*, que permite a modelagem eficiente de dependências de longo alcance entre palavras (Vaswani *et al.*, 2021).

Para que os modelos de linguagem possam processar automaticamente a linguagem natural, é necessário que o texto seja representado adequadamente. Nesse contexto, são dois os conceitos fundamentais: *tokens* e *embeddings*. Os *tokens* representam as menores unidades de texto utilizadas pelo modelo, enquanto os *embeddings* são representações vetoriais dessas unidades, o que preserva as propriedades semânticas e contextuais.

2.4.3.1 Tokens

A tokenização é o processo de segmentação do texto em unidades menores chamadas *tokens* que podem corresponder a palavras, subpalavras, caracteres ou até sentenças inteiras. Essa etapa é essencial para que os modelos de linguagem possam interpretar e processar dados textuais de maneira estruturada (Webster; Kit, 1992). As principais estratégias de tokenização incluem:

- **Tokenização por palavras**: Divide o texto em palavras individuais, sendo adequada para idiomas onde os espaços separam palavras.
- **Tokenização por subpalavras**: Segmenta palavras em unidades menores, facilitando o processamento de palavras raras e variações morfológicas.
- **Tokenização por caracteres**: Divide o texto em caracteres individuais, útil para sistemas que lidam com idiomas sem separação clara entre palavras.
- **Tokenização por sentenças**: Trata sentenças inteiras como unidades, relevante para tarefas como resumo automático e análise discursiva.

A escolha da estratégia de tokenização impacta diretamente o desempenho dos modelos de linguagem, influenciando sua capacidade de lidar com diferentes idiomas e estilos textuais (Petrov *et al.*, 2023).

2.4.3.2 Embeddings

Os *embeddings* são representações vetoriais de palavras ou frases que permitem que modelos de aprendizado de máquina capturem informações semânticas e contextuais da linguagem. Esses vetores são gerados a partir de grandes corpora textuais e possibilitam que palavras semanticamente similares tenham representações próximas no espaço vetorial. Entre as principais técnicas de geração de *embeddings*, destacam-se:

- **Word2Vec** (Mikolov *et al.*, 2013): Modelo baseado em redes neurais que aprende representações distribuídas das palavras.
- **GloVe** (Pennington; Socher; Manning, 2014): Método que utiliza matrizes de coocorrência para capturar relações globais entre palavras dentro de um corpus.
- **Transformer Embeddings** (Vaswani *et al.*, 2021): Representações contextuais extraídas de modelos baseados em *Transformers*, como BERT e GPT, que ajustam os *embeddings* conforme o contexto da sentença.

A utilização de *embeddings* aprimora significativamente a eficiência dos modelos de PLN, permitindo que compreendam melhor as relações semânticas e sintáticas entre palavras.

2.4.4 Grandes Modelos de Linguagem

Os LLMs são modelos de linguagem avançados, treinados em grandes conjuntos de dados textuais — muitas vezes na ordem de *terabytes* — o que lhes permite compreender e gerar textos com elevada precisão e naturalidade. Esses modelos são empregados em diversas aplicações, como assistentes virtuais, chatbots, motores de busca e ferramentas de escrita assistida (Bowman, 2023).

Uma característica central dos LLMs é sua capacidade de capturar nuances da linguagem humana, como ambiguidades, ironias e o contexto pragmático, o que torna a interação entre humanos e máquinas mais natural e eficaz. Isso é possível graças ao uso de arquiteturas como os *Transformers*, que permitem a modelagem eficiente de relações complexas entre palavras e frases (Peng *et al.*, 2023). Entre os LLM mais notáveis, destacam-se:

- **GPT (Generative Pre-trained Transformer)**: Desenvolvido pela OpenAI, inclui modelos como GPT-3.5 e GPT-4, amplamente utilizados em aplicações como o *ChatGPT* e o *Microsoft Copilot*.
- **GEMINI**: Desenvolvido pelo Google *DeepMind*, é uma família de modelos multimodais projetados para integrar texto, imagens e outros tipos de dados, com ênfase em tarefas complexas de raciocínio e geração de conteúdo.
- **Qwen**: Desenvolvido pela Alibaba Cloud, é uma série de modelos de código aberto com forte desempenho em tarefas multilíngues e aplicações em geração de texto, assistentes virtuais e tradução automática.
- **LLaMA (Large Language Model Meta AI)**: Desenvolvido pela Meta, sendo um modelo de código aberto voltado para pesquisas em PLN, com foco em eficiência e acessibilidade para a comunidade acadêmica.
- **Phy**: É uma família de modelos de linguagem voltada para pesquisas em física e ciências exatas, especializada na compreensão e geração de textos técnicos, resolução de problemas e suporte a pesquisadores.
- **Zephyr**: Trata-se de uma linha de modelos leves e otimizados para dispositivos com recursos limitados, buscando oferecer capacidades de compreensão e geração de linguagem natural em contextos embarcados e de borda.

A eficácia dos LLM decorre da combinação de redes neurais e grandes volumes de dados, permitindo-lhes gerar previsões precisas e produzir respostas contextualmente relevantes. Esses modelos têm sido importantes para avanços na inteligência artificial aplicada à linguagem natural.

2.4.5 *Prompt Engineering*

A engenharia de *prompt* é uma técnica essencial para otimizar a interação entre humanos e modelos de linguagem, com o objetivo de gerar respostas mais precisas e alinhadas às expectativas dos usuários. As técnicas envolvem a personalização da saída, o controle de contexto e a criação de interações dinâmicas, visando tornar as respostas mais eficazes e adaptáveis ao contexto específico (Brown *et al.*, 2020; Markowitz; Warkentin; Dyer, 2023).

De acordo com o artigo de Markowitz, Warkentin e Dyer (2023), foram identificados diferentes padrões de uso de *prompts*, que representam estratégias estruturadas para interagir de maneira eficaz com modelos de linguagem, permitindo adaptações conforme necessário. Entre os principais padrões estão:

- Customização da Saída: Definir o formato e estilo das respostas, como ao solicitar explicações simplificadas ou detalhadas para um público específico.
- Controle de Contexto: Estabelecer restrições sobre o escopo das respostas, garantindo foco em aspectos específicos da questão.
- Interação: Criar diálogos iterativos para refinar as respostas ao longo do tempo.

Além disso, a personalização dos modelos pode incluir ajustes para melhorar a acessibilidade, como adaptar *chatbots* para linguagem simplificada, oferecer suporte a múltiplos idiomas e garantir recursos de inclusão para pessoas com deficiência.

O conceito de *prompt engineering*, focado na estruturação de instruções de forma clara e eficiente, busca gerar respostas mais alinhadas às necessidades dos usuários ao interagirem com modelos de linguagem baseados em IA (Brown *et al.*, 2020). Esse campo de pesquisa visa otimizar a interação entre humanos e computadores, ajustando os *prompts* de acordo com o domínio de aplicação para aumentar a precisão, a relevância e a eficiência dos resultados.

Para cenários mais complexos, como a geração de planos de ação ou a iteração sobre respostas anteriores, os padrões de *prompt* também se mostram altamente eficazes. Um exemplo avançado seria:

- **Prompt**: “Imagine que você é um consultor de TI responsável por melhorar a infraestrutura de uma empresa que lida com grandes volumes de dados. Forneça um plano de ação detalhado para otimização de processos, considerando custos e recursos. Após apresentar o plano, reavalie suas sugestões com base em soluções alternativas.”

Essa abordagem de reavaliação e iteração ajuda a refinar o conteúdo gerado pelo modelo.

A especificidade no detalhamento dos *prompts* também é essencial para obter respostas mais precisas. Veja o exemplo a seguir, que ilustra como o nível de detalhe influencia a resposta:

- **Prompt vago**: “Explique inteligência artificial.”

- **Prompt detalhado**: “Explique o conceito de inteligência artificial, mencionando seus principais ramos e fornecendo exemplos práticos de aplicação.”

Fornecer maior detalhamento aumenta a probabilidade de o modelo compreender melhor a intenção do usuário e oferecer uma resposta mais rica e contextualizada.

Incluir exemplos específicos na formulação do *prompt* facilita a compreensão do formato esperado para a resposta. Por exemplo:

- **Prompt sem exemplo**: “Liste três benefícios da automação.”
- **Prompt com exemplo**: “Liste três benefícios da automação. Por exemplo, um benefício da automação na indústria é a redução de erros humanos.”

Essa técnica contribui para garantir respostas mais alinhadas com as expectativas do usuário.

Definir claramente a estrutura da resposta também é útil para guiar o modelo. Por exemplo, ao solicitar uma explicação, pode-se especificar a organização do conteúdo:

- **Prompt**: “Explique a teoria das redes neurais em três parágrafos, destacando conceito, funcionamento e aplicações.”

Essa definição ajuda a garantir coerência e organização nas respostas geradas.

A técnica de *role-playing* consiste em pedir ao modelo para assumir um papel específico, influenciando o tom e o nível de complexidade da resposta. Exemplo:

- **Prompt**: “Imagine que você é um professor de ciência da computação. Explique o conceito de aprendizado de máquina para iniciantes.”

Esse recurso permite ajustar o nível de detalhamento de acordo com o público-alvo.

O refinamento iterativo é outra técnica importante para aprimorar continuamente as respostas. Isso pode ser feito ajustando os *prompts* ao longo do processo, de modo a refinar a qualidade do conteúdo gerado.

Estabelecer restrições claras, como limites de palavras ou foco em aspectos específicos, também é uma estratégia eficaz. Por exemplo:

- **Prompt**: “Resuma o impacto da inteligência artificial na economia em até 100 palavras.”

Assim, é possivel definir a extensão da resposta ou outras restrições mantém o texto conciso e concentrado nas informações mais relevantes.

A *prompt engineering* oferece um conjunto de técnicas que podem ser combinadas para otimizar a interação com modelos de linguagem. A adaptação desses padrões aos diferentes domínios e uma documentação bem estruturada são fundamentais para explorar todo o potencial dessas ferramentas.

2.4.6 Técnicas de Fine-Tuning para adaptação de LLMs

O *fine-tuning* é uma etapa central no processo de adaptação de modelos de linguagem de grande porte LLMs a tarefas específicas. Trata-se da técnica de ajuste dos parâmetros de um modelo previamente treinado em grandes volumes de dados, utilizando um conjunto de dados menor e mais representativo de uma nova tarefa ou domínio de interesse (Devlin *et al.*, 2019; Howard; Ruder, 2018). Ao invés de treinar o modelo do zero, o *fine-tuning* reaproveita o conhecimento adquirido durante o treinamento original, tornando o processo mais eficiente em termos computacionais e mais eficaz na especialização para tarefas-alvo (Raffel *et al.*, 2020).

Essa abordagem tem sido amplamente utilizada em modelos como BERT (Devlin *et al.*, 2019), ULMFiT (Howard; Ruder, 2018), T5 (Raffel *et al.*, 2020), entre outros, demonstrando ganhos significativos em tarefas de interpretação de linguagem natural, classificação de texto e compreensão semântica. A seguir, são listadas algumas das principais técnicas de *fine-tuning* aplicáveis a LLMs:

- **Fine-tuning completo:** técnica tradicional em que todos os parâmetros do modelo são atualizados durante o ajuste à nova tarefa. Apesar de eficaz, pode ser computacionalmente custoso e menos viável para modelos muito grandes.

- **Fine-tuning parcial:** apenas parte dos parâmetros do modelo (geralmente das últimas camadas) é ajustada, reduzindo o custo computacional e mitigando o risco de *overfitting* (Howard; Ruder, 2018).

- **Uso de *Adapters*: pequenas camadas adicionais, chamadas de módulos intermediários leves, são inseridas entre as camadas do modelo original. Esses módulos contêm um número reduzido de parâmetros em comparação ao modelo completo e, portanto, tornam o treinamento mais eficiente. Apenas eles são ajustados durante o**

processo de adaptação, o que permite reutilizar o mesmo modelo base em múltiplas tarefas sem que haja interferência entre elas (Pfeiffer *et al.*, 2020). Entre as variações dessa técnica, destaca-se o LoRA *(Low-Rank Adaptation)*, que realiza ajustes eficientes por meio da decomposição de matrizes de peso, permitindo atualizações com menos parâmetros e mantendo o modelo base congelado (Hu *et al.*, 2021).

- **Instrução com *prompt tuning*: em vez de ajustar diretamente os parâmetros do modelo, essa técnica insere vetores de *prompt* aprendidos junto às entradas originais. Esses vetores funcionam como instruções adicionais que orientam o comportamento do modelo para tarefas específicas, alcançando efeitos semelhantes ao fine-tuning, mas sem necessidade de treinar todo o modelo (Gao; Fisch; Chen, 2021).**

Vale destacar que abordagens como *zero-shot*, *one-shot*, *transfer learning* e *meta-learning* são paradigmas distintos que dizem respeito à capacidade dos modelos de generalizarem conhecimento para novas tarefas, mas não constituem, conceitualmente, técnicas de *fine-tuning*. Por exemplo:

- ***Zero-shot learning***: o modelo aplica conhecimento prévio a tarefas para as quais não foi treinado explicitamente, sem qualquer ajuste de parâmetros. Exemplo: um modelo treinado em um corpus amplo é capaz de redigir textos sobre um novo domínio sem exemplos específicos prévios (Zaremba; Sutskever, 2014; Bowman, 2023; Markowitz; Warkentin; Dyer, 2023).

- ***One-shot learning***: o modelo realiza uma tarefa com base em um único exemplo de treinamento. Requer forte capacidade de generalização, mas não envolve o ajuste dos parâmetros do modelo como no *fine-tuning*. Exemplo: traduzir frases entre idiomas a partir de um único exemplo de tradução (Bowman, 2023; Markowitz; Warkentin; Dyer, 2023).

- ***Transfer learning*: refere-se ao aproveitamento de conhecimento adquirido em uma tarefa ou domínio mais amplo para melhorar o desempenho em outra tarefa, geralmente mais específica, mas relacionada. Em vez de treinar um modelo do zero, parte-se de um modelo previamente treinado em grandes volumes de dados, o que reduz custo computacional e tempo de treinamento (Mikolov *et al.*, 2013; Bowman, 2023; Markowitz; Warkentin; Dyer, 2023).**

O *fine-tuning* pode ser entendido como uma etapa dentro do *transfer learning*, na qual os parâmetros do modelo pré-treinado são ajustados com dados de uma tarefa-alvo. Embora relacionados, os termos não são sinônimos: *transfer learning* descreve o conceito mais geral de reutilização de conhecimento, enquanto o *fine-tuning* é uma técnica prática que implementa essa ideia.

Exemplo: um modelo treinado para geração de texto pode ser posteriormente ajustado, por meio de *fine-tuning*, para uma aplicação específica como análise de sentimentos.

- ***Meta-learning***: propõe que o modelo aprenda a aprender, ou seja, seja capaz de se adaptar rapidamente a novas tarefas com poucos exemplos, sem necessariamente requerer ajustes intensivos de seus parâmetros. Exemplo: adaptar-se rapidamente à tradução de um novo idioma a partir de poucos exemplos, sem treinamento completo (Vaswani *et al.*, 2021; Bowman, 2023; Markowitz; Warkentin; Dyer, 2023).

Essas abordagens podem ser complementares ao *fine-tuning*, mas devem ser entendidas como estratégias distintas no campo do aprendizado de máquina, especialmente no contexto da adaptação de modelos de linguagem de grande porte.

2.5 *Chatbots*

Os *Chatbots* são sistemas que empregam algoritmos de PLN para interpretar e responder às consultas dos usuários utilizando uma linguagem natural. Esses programas, baseados em IA, simulam conversas humanas, permitindo interações com dispositivos digitais que se assemelham a diálogos com pessoas reais (Turing, 1950).

Esses assistentes virtuais e sistemas automáticos de resposta são desenvolvidos com o intuito de tornar a interação mais intuitiva e eficaz, proporcionando uma experiência mais próxima da comunicação humana. Ao integrar as técnicas de PLN nessas tecnologias, é possível aprimorar a capacidade de compreensão contextual, permitindo respostas mais precisas, relevantes e personalizáveis de acordo com um determinado contexto.

O primeiro *chat* foi desenvolvido entre 1964 e 1966 pelo cientista da computação germano-americano Joseph Weizenbaum, o *chat* ficou conhecido como ELIZA. Esse *chat* é reconhecido por sua habilidade em enganar humanos, fazendo-os acreditar que estão conversando com outro ser humano. ELIZA foi criado de forma não comercial, com o objetivo de satirizar as respostas

de um psicoterapeuta em uma entrevista psiquiátrica inicial. Seguindo a estrutura adotada por muitos chatbots modernos, ELIZA pré-determina respostas de texto para serem acionadas por entradas específicas, antecipando as interações do usuário (Zemčík, 2019).

Segundo Muldowney (2017), os *chatbots* são fundamentais para diversas aplicações, incluindo atendimento ao cliente, suporte técnico, *marketing*, vendas, entre outros. Existem dois tipos de *chatbots*, os declarativos e os conversacionais. Os *chatbots* declarativos, ou orientados a tarefas, são *chats* que executam funções específicas utilizando regras previamente introduzidas, ou seja, possui seu processo de aprendizado pré-definido e limitado. Já os *chatbots* conversacionais são orientados à predição, isto é, são programados para antecipar as necessidades e intenções dos usuários com base no contexto da conversa e nas interações anteriores. Eles são assistentes virtuais mais sofisticados, interativos e personalizados, assim, possuem maior capacidade de compreensão e de aprendizado contextual (Dale, 2016). Atualmente muitos *chatbots* são alimentados com LLMs para obter um melhor desempenho.

2.5.1 Desenvolvimento de um *Chatbot*

Os *chatbots* são desenvolvidos por meio de um treinamento que envolve uma variedade de técnicas e modelos de linguagem. Com base em uma análise abrangente dos estudos de Hussain, Sianaki e Ababneh (2019), Maglogiannis, Iliadis e Pimenidis (2020) e Choudhary e Chauhan (2023), é possível identificar as etapas comuns no treinamento e desenvolvimento desses sistemas. Abaixo, são descritas essas etapas:

1. Coleta de dados: os *chatbots* podem ser treinados usando conjuntos de dados de conversas humanas reais. Esses dados podem ser extraídos de fontes como conversas de bate-papo, redes sociais, fóruns *online*, *e-mails*, entre outros. Quanto mais variados e representativos forem os dados, melhor será o desempenho do *chatbot*.

2. **Pré-processamento de dados**: antes de utilizar os dados no treinamento do *chatbot*, podem ser necessárias etapas de pré-processamento, como tokenização, remoção de pontuação, padronização para letras maiúsculas ou minúsculas, remoção de stop words (palavras muito frequentes que pouco contribuem para o significado), lematização (redução de palavras flexionadas ao seu lema), entre outras técnicas.

3. Escolha do modelo de linguagem: os modelos de linguagem são a base dos *chatbots*. Esses modelos são treinados em grandes conjuntos de dados textuais para prever a próxima palavra dado uma sequência de palavras.

4. Treinamento do modelo de linguagem: o modelo de linguagem escolhido é treinado nos dados pré-processados. Durante esse treinamento, o modelo ajusta seus parâmetros para maximizar a probabilidade de prever corretamente a próxima palavra, dada uma sequência de palavras. Isso é feito iterativamente usando técnicas de otimização, como por exemplo a descida do gradiente estocástico.

5. Ajuste fino: em alguns casos, os *chatbots* podem passar por uma etapa de ajuste fino, na qual o modelo de linguagem pré-treinado é ajustado para um conjunto específico de dados, o que permite ele a realizar uma tarefa específica naquele contexto ajustado. Isso pode ajudar a melhorar o desempenho do *chatbot* em um cenário específico.

6. Construção da interface do *chatbot*: uma vez que o modelo de linguagem é treinado, ele pode ser integrado a uma interface de *chat*, como um aplicativo móvel, *site* ou plataforma de mensagens. A interface permite que os usuários interajam com o *chatbot* ao enviar mensagens de texto, fornecendo assim uma aplicação.

7. Avaliação e iteração: após a disponibilização do *chatbot*, é importante avaliar continuamente seu desempenho e coletar principalmente *feedbacks* dos usuários. Assim, com base nessas informações, o *chatbot* pode ser aprimorado e refinado para oferecer uma experiência de usuário mais satisfatória.

2.5.2 *Chatbots* e Acessibilidade

A existência de *chats* que sejam acessíveis é um ponto primordial para garantir uma comunicação inclusiva e eficiente para todos os usuários, independentemente de suas habilidades ou necessidades específicas. Os *chatbots* estão progressivamente ganhando espaço em diversos serviços, sejam eles no setor privado ou no setor público (Hussain; Sianaki; Ababneh, 2019; Moreira; Pazoti; Siscoutto, 2022).

Os requisitos de acessibilidade são uma demanda de exigência primitiva, pois torna principalmente a comunicação mais inclusiva (Libraria, 2023.). Alguns exemplos de condições

acessíveis são: suporte a leitores de tela, fontes legíveis, opções de contraste, interfaces adaptáveis e a forma de entrega e recebimento de informações pelas aplicações (Costa *et al.*, 2023; Digital, 2024).

Empresas como *Microsoft* e *Google* já possuem a visão de uma comunicação voltada para a acessibilidade e apresentam iniciativas desde a etapa de contratação de seu pessoal, até em fornecer um *designer* inclusivo em seus produtos e aplicações (Microsoft, 2024.; Google, 2024.). Entretanto, a acessibilidade voltada para *chatbots*, ainda é um campo a ser explorado.

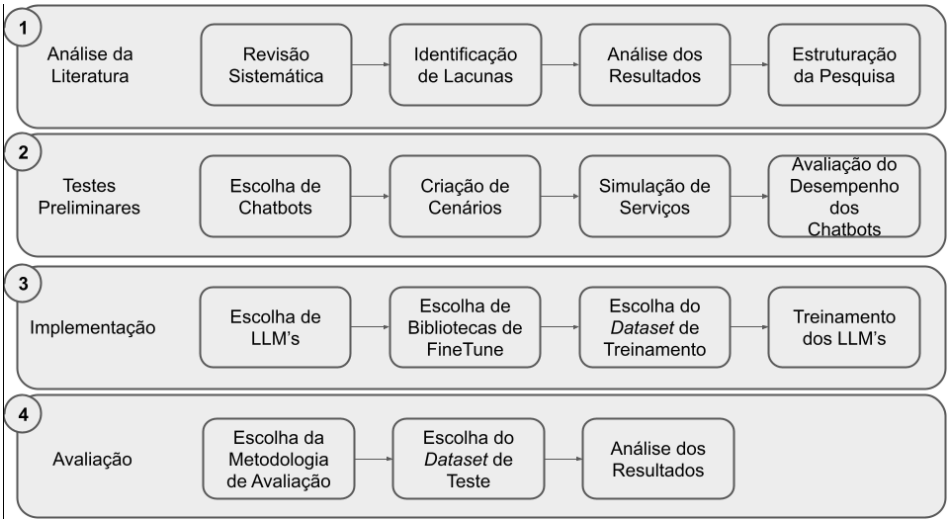
Assim, ao estudar e desenvolver *chatbots* acessíveis, as empresas e desenvolvedores não apenas colaboram com as exigências legais de acessibilidade, mas também proporcionam uma experiência mais abrangente, com a visão de inclusão, para que todos os usuários possam participar ativamente no uso dessas tecnologias e aproveitar os benefícios oferecidos por esses sistemas, de forma igualitária.

3 METODOLOGIA

Este capítulo apresenta a abordagem metodológica adotada para o desenvolvimento desta pesquisa, cujo objetivo principal é avaliar a viabilidade do uso de LLMs para interpretar textos escritos na variante GLOSA, contribuindo para o desenvolvimento de chatbots mais acessíveis à comunidade surda, que utiliza o Português como segunda língua. Para isso, foram definidos quatro objetivos específicos: (1) avaliar soluções de chatbots atualmente disponíveis, analisando sua familiaridade com o modo de escrita dos usuários surdos; (2) verificar a capacidade de interpretação de textos em GLOSA por modelos de linguagem sem aplicação de *fine-tuning*; (3) investigar o impacto do *fine-tuning* em diferentes configurações na adaptação dos modelos à variante GLOSA; e (4) comparar os resultados obtidos entre diferentes modelos e estratégias, a fim de propor recomendações para o uso em sistemas conversacionais voltados a esse público.

As etapas da pesquisa foram organizadas de forma sequencial para garantir uma análise consistente, contemplando: a revisão da literatura, a experimentação preliminar com chatbots bancários, a implementação de modelos de linguagem ajustados à GLOSA e, por fim, a avaliação dos resultados obtidos. A Figura 3.1 apresenta uma visão geral das etapas metodológicas desta investigação pesquisa.

Figura 3.1 – *Overview* das Etapas da Pesquisa



Fonte: Autora (2025).

1. Etapa 1 – Análise da Literatura

A primeira etapa consistiu na realização de uma revisão sistemática da literatura com foco em acessibilidade digital, modelos de linguagem e a interação de pessoas surdas com sistemas conversacionais automatizados. Foram consultadas bases acadêmicas de alto impacto, como IEEE Xplore, ACM Digital Library, Scopus e Web of Science, com critérios definidos para busca, inclusão e exclusão.

Essa revisão permitiu a identificação de lacunas (*gaps*) na literatura, revelando limitações recorrentes na forma como *chatbots* lidam com textos escritos em Português como segunda língua. A análise dos trabalhos levantados serviu de base para a formulação dos objetivos específicos desta pesquisa, bem como para a definição das diretrizes metodológicas adotadas nas etapas seguintes. Os procedimentos detalhados desta fase encontram-se descritos no Capítulo 4.

2. Etapa 2 – Testes Preliminares

   Com base nas lacunas identificadas na literatura, a segunda etapa teve como objetivo avaliar a eficácia de *chatbots* bancários amplamente utilizados no Brasil quanto à sua capacidade de compreender a variante GLOSA. Para isso, foram selecionados os assistentes virtuais BIA (Bradesco), o chatbot do Banco do Brasil e o do Itaú, por serem amplamente utilizados e aplicarem técnicas modernas de Inteligência Artificial em sua operação (IBM, 2019)(Associação dos Gerentes do Banco do Brasil, 2022) (Conteúdo, 2024).

   A interação com os *chatbots* ocorreu por meio do *WhatsApp* e foi estruturada com base nos seguintes procedimentos:

   - **Criação de Cenários**: simulação de interações bancárias típicas, incluindo pagamentos, transferências e consultas de saldo.
   - **Simulação de Serviços**: registro das conversas realizadas com foco na interpretação das mensagens em GLOSA e na adequação das respostas.
   - **Avaliação dos Desempenhos dos *Chatbots***: avaliação da precisão interpretativa, coerência das respostas e adequação da interface conversacional às necessidades de usuários surdos.

Este estudo preliminar evidenciou limitações na capacidade dos sistemas analisados em interpretar construções linguísticas típicas da GLOSA, o que reforçou a necessidade de

adaptação de modelos por meio de técnicas específicas. Os detalhes completos desta etapa estão descritos no Capítulo 5.

3. Etapa 3 – Implementação

   A terceira etapa compreendeu o desenvolvimento e o treinamento de modelos de linguagem capazes de interpretar textos na variante GLOSA. O processo envolveu as seguintes ações:

   - **Escolha de LLMs e Bibliotecas**: escolha de arquiteturas de *Large Language Models* e bibliotecas compatíveis com técnicas de *fine-tuning*, com base em requisitos de performance e escalabilidade.
   - **Escolha de *Dataset* de Treinamento**: construção e organização de um conjunto de dados contendo frases e sentenças representativas da variante GLOSA, a partir de traduções figurativas e exemplos reais de uso.
   - **Treinamento dos LLMs**: foi realizado o *fine-tuning* utilizando parâmetros base e configurações personalizadas via API, como ajuste de temperatura, controle de repetição e número de épocas. Como exemplo, o modelo *Gemini* 1.5 Flash foi empregado na tarefa de tradução do Português para a GLOSA, com o objetivo de facilitar a comunicação acessível e apoiar pesquisas voltadas à inclusão de pessoas surdas usuárias dessa forma de representação linguística.

   Essa etapa está descrita com profundidade no Capítulo 6, incluindo a fundamentação técnica para as escolhas realizadas.

4. Etapa 4 – Avaliação

   A última etapa metodológica buscou avaliar a eficácia dos modelos ajustados na tarefa de interpretação da GLOSA, por meio da definição de métricas, conjuntos de dados de teste e procedimentos de análise.

   Essa avaliação foi organizada em três componentes principais:

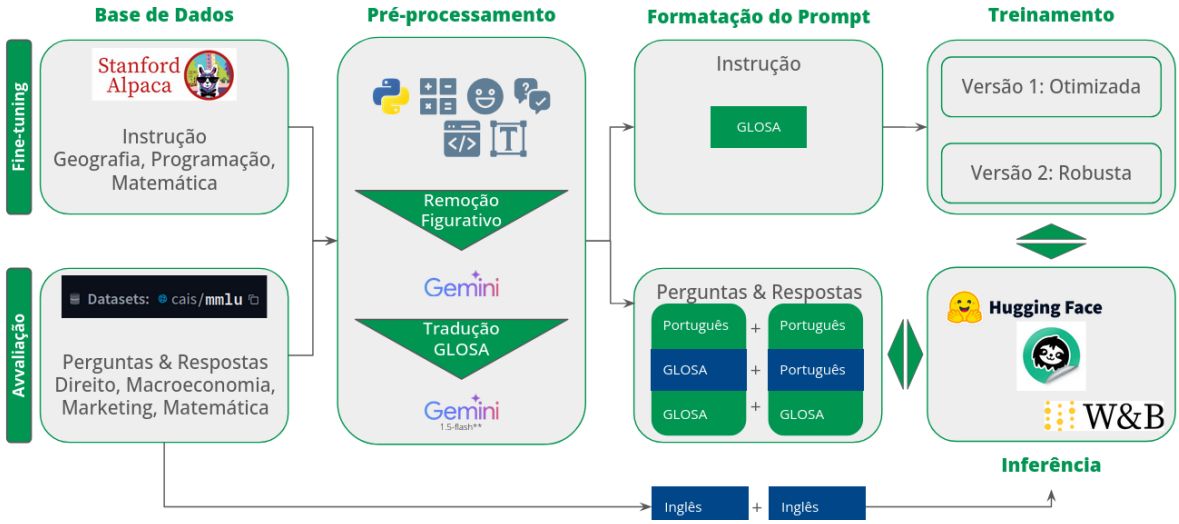
   - **Escolha da Metodologia de Avaliação**: aplicação de métricas objetivas (como acurácia) e subjetivas (como avaliação humana da qualidade das respostas).
   - **Seleção do Dataset de Teste**: utilização de dados representativos de situações de comunicação cotidiana em GLOSA.

- **Análise dos Resultados**: comparação do desempenho entre diferentes modelos e configurações, de modo a fornecer subsídios para recomendações práticas e futuras pesquisas.

Os resultados obtidos, bem como a discussão sobre os impactos da abordagem adotada, estão apresentados no Capítulo 7.

As Etapas 3 e 4, correspondentes à implementação e à avaliação dos modelos, estão representadas de forma sintética no fluxo macro ilustrado na Figura 3.2. A figura não descreve minuciosamente cada procedimento técnico, mas oferece uma visão geral da arquitetura metodológica empregada, evidenciando a articulação entre ferramentas, tecnologias e processos. Assim, torna-se possível compreender o encadeamento entre o treinamento dos modelos e a etapa de avaliação, situando ambas no contexto mais amplo da pesquisa. Os detalhes específicos de cada etapa são discutidos em capítulos posteriores.

Figura 3.2 – *Overview Fine-tuning* e Avaliação



Fonte: Autora (2025).

A metodologia delineada neste capítulo busca integrar os aspectos científicos e inovações tecnológica na construção de soluções conversacionais mais inclusivas para a comunidade surda. Ao seguir uma sequência estruturada — análise da literatura, testes exploratórios, implementação técnica e avaliação crítica —, a pesquisa estabelece um referencial metodológico para futuras investigações na área de acessibilidade digital e processamento de linguagem natural.

4 REVISÃO DA LITERATURA

Segundo Booth, Colomb e Williams (2008), a pesquisa bibliográfica é uma prática que envolve a identificação, coleta e análise de informações relevantes disponíveis na literatura acadêmica. Ela oferece uma base sólida para a construção do conhecimento ao fornecer percepções, referências e contextos necessários para embasar investigações subsequentes.

O processo de investigação da literatura busca reunir e analisar as informações disponíveis em fontes bibliográficas tais como: livros, artigos científicos, teses, dissertações, entre outras publicações acadêmicas. Essa etapa permite coletar informações para situar o trabalho dentro do contexto existente das literaturas já consolidadas. Assim, é possível identificar as lacunas pré-existentes na literatura, reconhecer os debates em curso e construir uma base teórica sólida para o desenvolvimento de uma pesquisa (Flick, 2018).

4.1 Questões Chaves

Ao conduzir uma pesquisa na literatura é essencial estabelecer objetivos claros, muitas vezes expressos na forma de questões a serem respondidas em relação ao tema investigado (Booth; Colomb; Williams, 2008). Essas questões-chaves fornecem uma estrutura para orientar a revisão da literatura e direcionar a análise dos resultados. No contexto deste estudo, foram delineadas as seguintes indagações a serem exploradas:

1. Quais são as preocupações com a inclusão do público alvo e os principais trabalhos desenvolvidos envolvendo o uso de tecnologia?
2. Quais são as principais técnicas utilizadas no desenvolvimento de *chatbots* ou LLMs?
3. Como é testado e avaliado o desempenho dessas tecnologias na comunidade surda?
4. Existe algum trabalho voltado para essa área envolvendo o uso de *chatbots* ou LLMs?

4.2 Critérios de Aceitação

Definir critérios de aceitação e rejeição é interessante para o processo de revisão e seleção de trabalhos acadêmicos, uma vez que esses critérios garantem a padronização e imparcialidade nas avaliações de tal modo a garantir a qualidade e relevância dos trabalhos selecionados.

Além disso, esses critérios fornecem orientação clara para embasamento de novos trabalhos, ao direcionar os esforços para atender aos requisitos já esperados. Assim, os critérios ajudam a garantir qualidade no processo de revisão e contribuem para o avanço do conhecimento científico. Para este trabalho, os critérios de aceitação e rejeição são apresentados no Quadro 4.1.

Quadro 4.1 – Critérios de Avaliação dos Estudos.

Aceitação	**Rejeição**
A escrita do trabalho deve ser em Português ou Inglês	Não ter o uso de tecnologia envolvido
Abordar temas de linguagem escrita e não gestual	Não ter a metodologia clara e definida para reprodução de experimento quando utilizado datasets públicos
Utilizar tecnologia como meio de acessibilidade	Abordar reconhecimento de gestos

Fonte: Autora (2024).

4.3 Repositórios de Literatura

Após a elaboração do escopo em que serão selecionados os estudos para análise, é necessário também definir onde serão coletados os estudos para revisão da literatura. Existem várias ferramentas de busca disponíveis para encontrar artigos e estudos acadêmicos relevantes, algumas das mais populares e amplamente utilizadas incluem:

- ACM [1]: é uma das principais organizações dedicadas à computação, ela oferece uma ampla gama de recursos para profissionais e pesquisadores da área, incluindo uma biblioteca digital abrangente que inclui diversos tópicos em ciência da computação.
- Google Acadêmico[2]: é uma plataforma de pesquisa que indexa uma ampla gama de literatura acadêmica em diversas disciplinas, incluindo artigos, trabalhos de conclusão de cursos, monografias, teses, entre outros.
- IEEE Xplore [3]: é uma biblioteca digital que oferece acesso a artigos, conferências e padrões relacionados a tecnologia, engenharia e ciência da computação.

[1] *Site* oficial ACM: https://dl.acm.org/
[2] *Site* oficial Google Acadêmico: https://scholar.google.com
[3] *Site* oficial IEEE Xplore: https://ieeexplore.ieee.org

- Periódicos CAPES [4]: referem-se a um conjunto de revistas científicas e acadêmicas disponibilizadas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Essa iniciativa proporciona acesso gratuito e online a milhares de periódicos em diversas áreas do conhecimento, como ciências sociais, saúde, humanidades, exatas e biológicas.

4.4 *Strings* de Busca

As plataformas de pesquisa permitem aos usuários realizar buscas utilizando palavras-chaves específicas ou sequências de caracteres, conhecidas como "*string de busca*". Esse termo se refere a uma sequência de caracteres que pode incluir operadores lógicos para indicar a relação entre os termos dessa sequência, possibilitando assim uma pesquisa mais precisa e refinada.

A fim de realizar uma investigação abrangente e detalhada das áreas e tópicos pertinentes ao tema em questão, foram definidas e elaboradas *strings* de buscas específicas para fins de simplificação. Além disso, as *strings* foram buscadas tanto em Português quanto em inglês, sendo que em inglês, foi incluído o termo "ASL", que é a LS dos Estados Unidos da América e do Canadá (Bahan, 1996).

Este estudo aborda duas áreas distintas, porém, interligadas, que desempenham papeis importantes no cenário atual da comunicação: a comunicação realizada por surdos, com foco em indivíduos para os quais o Português é uma L2, e o campo do processamento de linguagem natural, com ênfase nas tecnologias relacionadas à comunicação escrita, os *chatbots*.

A interseção desses dois domínios de estudo reside no desenvolvimento e implementação de tecnologias acessíveis, especialmente no contexto de *chatbots*, para atender às necessidades específicas do público que utiliza o Português como L2. Dessa forma, as *strings* de busca foram elaboradas de maneira a capturar efetivamente essas áreas temáticas e suas interconexões, visando garantir uma abordagem abrangente na coleta de dados e na análise subsequente.

As buscas foram realizadas em três momentos distintos. O primeiro correspondeu ao período inicial do trabalho, em que foram coletados estudos para servir de insumo à pesquisa (2023). O segundo ocorreu após o refinamento do projeto, visando consolidar os projetos já encontrados (2024). Por fim, o terceiro momento deu-se na fase de consolidação, a fim de

[4] *Site* oficial Periódico CAPES https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/

verificar se houve alterações ou o surgimento de novos trabalhos que pudessem agregar valor ao trabalho desenvolvido (2025).

O Quadro 4.2 apresenta os termos buscados para o contexto de Libras e Português como L2, bem como o ASL e o Inglês como segunda língua.

Quadro 4.2 – Surdez, Comunicação e Linguagem de Sinais.

Id	**Português**	**Inglês**
SBLS-01	(“LIBRAS” OR “linguagem de sinais” OR “segunda língua” OR “comunidade surda”)	(“ASL” OR “sign language” OR “second language” OR “deaf community”)
SBLS-02	(“reconhecimento de linguagem de Sinais” OU “intérprete de sinais”)	(“sign language recognition” OR “sign language interpreter”)

Fonte: Autora (2024).

- id SBLS-01: o escopo da busca é amplo, e abrange uma variedade de termos relacionados à linguagem de sinais e à comunidade surda em ambos idiomas, Português e Inglês.
- id SBLS-02: foca especificamente em termos relacionados ao reconhecimento e interpretação da linguagem de sinais.

Já o Quadro 4.3 faz referência aos termos de *chatbots*, PLN e IA. Abaixo é definido o escopo de abrangência de cada *string* de acordo com seu identificador (id), associado:

Quadro 4.3 – *Chatbots* e Inteligência Artificial.

Id	**Português**	**Inglês**
SBIA-01	“arquiteturas de *chatbot* conversacional” AND “modelos de *chatbot*” AND “*frameworks* de *chatbot*” AND “tecnologias para *chatbots*”	“*conversational chatbot architectures*” AND “*chatbot models*” AND “*chatbot frameworks*” AND “*chatbot technologies*”
SBIA-02	(“processamento de linguagem natural” OR “PLN”) AND “*chatbots*”	(“*natural language processing*” OR “NLP”) AND “*chatbots*”

Fonte: Autora (2024).

- id SBIA-01: procura informações detalhadas sobre o desenvolvimento de *chatbots* conversacionais, evidenciando um alto nível de detalhamento ao explorar aspectos específicos do tema.

- id SBIA-02: possui um nível de detalhamento moderado, pois aborda os temas principais temas sem entrar em detalhes específicos.

Após refinados os trabalhos e contextualizado a respeito do tema, foi possível aprimorar as *strings* de busca, para trazer trabalhos precisos e concisos. Essas *strings* são observadas no Quadro 4.4, que são definidas abaixo, em seu respectivo escopo de abrangência, conforme seu identificador, explorado abaixo:

Quadro 4.4 – Interseção dos Campos.

Id	**Português**	**Inglês**
SB-01	(“*chatbot*” OR “agente de conversação” OR “assistente virtual”) AND (“Português L2” OR “língua portuguesa”) AND (“deficiência auditiva” OR “perda auditiva” OR “surdez”) AND (“linguagem natural” OR “linguagem de sinais” OR “linguagem oral” OR “tradução de GLOSA” OR GLOSA) AND (“acessibilidade” OR “design inclusivo” OR “tecnologia assistiva” OR “tecnologias”)	(“*chatbot*” OR “conversational agent” OR “virtual assistant”) AND (“English L2” OR “English language”) AND (“hearing impairment” OR “hearing loss” OR “deafness”) AND (“natural language” OR “sign language” OR “oral language” OR “translation of gloss” OR “GLOSS”) AND (“accessibility” OR “inclusive design” OR “assistive technology” OR “technologies”)
SB-02	(“LIBRAS” OR “linguagem de sinais” OR “segunda língua” OR “comunidade surda” OR “interação humano computador” OR “IHC” OR “processamento de linguagem natural” OR “NLP” OR “aprendizado de máquina”) NOT (“reconhecimento de linguagem de sinais” OR “interpretador de linguagem de sinais”)	(“ASL” OR “sign language” OR “second language” OR “deaf community” OR “human computer interaction” OR “HCI” OR “natural language processing” OR “NLP” OR “machine learning”) NOT (“sign language recognition” OR “sign language interpreter”)

Fonte: Autora (2024).

- id SB-01: aborda detalhes específicos sobre *chatbots* para usuários de língua portuguesa como segunda língua, com atenção especial para pessoas com deficiência auditiva. A busca é altamente detalhada, abordando vários aspectos, incluindo linguagem, acessibilidade e tecnologia assistiva. A complexidade é alta devido à diversidade de termos e à natureza técnica dos conceitos envolvidos.

- id SB-02: abrange uma variedade de tópicos, incluindo Libras, linguagem de sinais, segunda língua, comunidade surda, interação humano-computador, PLN e aprendizado de máquina. O nível de granularidade é alto, pois cobre uma diversidade de conceitos, embora a especificidade seja moderada, concentrando-se em áreas específicas dentro desses campos de estudo.

Os resultados das buscas para as *strings* presentes no Quadro 4.4 estão registrados no Quadro 4.5. Para refinar a seleção dos artigos mais relevantes para este projeto, foi adotada uma abordagem que segue uma sequência pré-definida de passos, os quais são delineados abaixo:

- Inicialmente, a busca foi conduzida na plataforma do Google Acadêmico, sendo esta a primeira base de dados consultada. Em seguida, foram exploradas outras fontes, culminando na busca na base de dados ACM, que representou o último ponto de pesquisa, conforme apresentado no Quadro 4.5.

- A análise dos artigos foi realizada considerando até a terceira página de resultados em cada plataforma. Nessa etapa, os artigos foram inicialmente avaliados com base em seus títulos. Aqueles que não estavam alinhados com as strings propostas para o trabalho foram descartados.

- Se, após a análise dos resultados até a terceira página, não fosse encontrada uma quantidade satisfatória de artigos relevantes, a busca era encerrada nessa etapa.

- Em seguida, os artigos selecionados tiveram seus resumos analisados. E se o resumo estivesse coerente com as *strings* de busca, esse artigo era selecionado para uma análise completa e detalhada.

- Os artigos selecionados passaram por uma leitura completa, a fim de verificar sua contribuição efetiva para o conteúdo da pesquisa em questão.

Quadro 4.5 – Resultados da busca da interseção dos campos.

		SB-01		**SB-02**	
		Português	**Inglês**	**Português**	**Inglês**
Google Acadêmico	Retornado	43	687	2.101	58.402
	Escolhido	5	3	3	5
Periódico CAPES	Retornado	264.720	215.007	0	1
	Escolhido	1	0	0	0
IEEE Xplore	Retornado	10.250	381.950	2	2
	Escolhido	0	2	0	0
ACM Digital Library	Retornado	236	117.512	3	48.955
	Escolhido	1	1	0	0
Total escolhido		7	6	3	5
Total revisado		21			

Fonte: Autora (2025).

Após a análise dos artigos e trabalhos selecionados para revisão, tornou-se evidente que a demanda por soluções acessíveis está em constante ascensão. No entanto, foi observado que esta necessidade ainda não foi plenamente abordada pela pesquisa existente, sugerindo uma lacuna significativa de estudos na área. Ao identificar este desafio, torna-se perceptível não apenas a ausência de uma aplicação dedicada ao público-alvo do projeto, mas também a complexidade do assunto abordado. Vale ressaltar que, ao realizar buscas em alguns repositórios, foram enfrentadas dificuldades devido à especificidade da *string* de busca, o que pode ter limitado a amplitude da pesquisa. Essa limitação não apenas evidencia a escassez de trabalhos relacionados a essa temática, mas também destaca a complexidade do cenário. Esta constatação reforça a necessidade de uma abordagem que não apenas atenda à carência de estudos, mas também considere a diversidade e complexidade do tema em foco.

Além disso, a transcrição e formulação de "GLOSAS" provindas das Linguagens Orais, para um idioma de texto apresenta-se como um desafio complexo e multifacetado, que carece de soluções consolidas na literatura. Paralelamente, os avanços tecnológicos e os modelos de linguagem exigem grandes volumes de dados para serem treinados, o que impõe um obstáculo adicional nesta jornada.

No entanto, a ideia da revisão bibliográfica foi preencher essa lacuna de tal modo a identificar e explorar os estudos mais relevantes e similares em ambos os contextos. A partir dessa

análise espera-se não apenas contribuir para a compreensão desses desafios interdisciplinares, mas também realizar a implementação de soluções neste campo emergente e promissor. Este processo de seleção resultou na identificação de oito trabalhos que se mostraram mais próximos dos temas e metodologias abordados nesta pesquisa. Portanto, esses artigos selecionados constituem a base empírica sobre a qual este trabalho se fundamenta, e são apresentados na Seção 4.5.

4.5 Trabalhos Relacionados

O trabalho dos autores Moreira, Pazoti e Siscoutto (2022) apresenta uma ferramenta desenvolvida para facilitar a criação de *chatbots* personalizados na área educativa. A metodologia envolve a implementação de modelos de linguagens baseados em recuperação de informação e generativos, os quais utilizam algumas técnicas de PLN como por exemplo, as redes neurais LSTM bidirecionais e um esquema de tradução automática sequência-para-sequência.

O modelo baseado em recuperação da informação demonstrou ser uma ferramenta que pode ser utilizada na área educativa, e permite que seus próprios usuários definam seus modelos para instruir sobre tópicos específicos em um formato de *chatbot* personalizado. As técnicas utilizadas pelos autores também serão utilizadas neste trabalho, para poder desenvolver um *chatbot* similar, baseado em recuperação da informação, adicionando dados de um contexto específico.

O estudo de Oliveira, Lima e Araujo (2015), no contexto de GLOSAS, apresenta uma solução computacional inovadora para tradução automática e eficiente de textos do Português brasileiro para a Libras, integrada ao VLIBRAS. A metodologia incorpora estratégias de simplificação textual e classificação morfossintática. Os testes realizados com 20 usuários surdos demonstraram uma aceitação positiva da solução, com 90% manifestando disposição para utilizá-la na tradução de textos ou vídeos, embora 35% tenham identificado alguns problemas gramaticais. Destaca-se a importância do VLIBRAS, um projeto do Laboratório de Aplicações de Vídeo Digital (LAViD) da Universidade Federal da Paraíba (UFPB) que integra GLOSAS e animações, de modo a tornar os conteúdos acessíveis em Libras. O estudo evidencia a necessidade de aprimorar a qualidade da tradução, explorando estratégias para lidar com aspectos sintático-semânticos e pragmáticos em futuras pesquisas. A semelhança com este projeto envolve a metodologia utilizada, que envolve o trabalho conjunto com intérpretes.

Já os autores Costa *et al.* (2023) abordam a implementação da acessibilidade para o público que possui alguma limitação, no contexto da TV 3.0. Eles focaram seus estudos em representação de legendas, língua de sinais e áudio descrição. Foram desenvolvidos alguns módulos para essas áreas, permitindo a personalização do conteúdo para diferentes usuários. Desafios, como a sincronização entre dispositivos e a velocidade de exibição da língua de sinais, foram alguns aspectos levantados pelo estudo. Por fim, o propósito do trabalho é tornar a TV 3.0 mais acessível, beneficiando pessoas com deficiência visual ou auditiva no contexto brasileiro. A solução proposta neste trabalho vem como um adicional de tecnologias no campo da acessibilidade, contribuindo de forma positiva para esse nicho de pesquisa.

No campo internacional, os autores Joksimoski *et al.* (2022) fizeram uma revisão abrangente da literatura, onde examinaram e analisaram uma série de estudos e pesquisas realizadas na área de linguagem de sinais. Eles identificaram diferentes abordagens, técnicas e tecnologias utilizadas, bem como suas aplicações e resultados. Os principais resultados incluem a constatação de que algoritmos de aprendizado de máquina e *deep leaning* podem melhorar o reconhecimento e interpretação da linguagem de sinais.

A maioria dos estudos revisados focalizou o uso de dados de entrada em formato de vídeo ou imagens, havendo também relatos sobre o emprego de dispositivos como luvas para a captura de gestos. Apesar da diversidade de línguas de sinais abordadas, não foram identificados estudos específicos sobre a língua de sinais brasileira. As análises englobaram o período de 2010 a 2021. Além disso, há uma expectativa de que avanços contínuos em *hardware* e abordagens de aprendizado de máquina levarão a melhorias significativas em todas as tecnologias relacionadas à linguagem de sinais, potencialmente permitindo a tradução em tempo real entre diferentes línguas de sinais e línguas faladas (Joksimoski *et al.*, 2022).

O estudo conduzido pelos autores (Khennouche *et al.*, 2023) oferece uma análise abrangente sobre a implementação e os desafios enfrentados no emprego do *ChatGPT* e *chatbots* generativos em Sistemas de Perguntas Frequentes (FAQs). A pesquisa explora as vastas possibilidades dessas tecnologias, nas quais se destacam a aplicabilidade em diversos setores como por exemplo, educação, finanças e atendimento ao cliente. Além disso, identifica limitações e questões éticas inerentes ao seu uso e analisa criticamente as métricas de avaliação existentes, no qual aponta inadequações na avaliação de *chatbots*.

Os autores também discutem o uso de modelos generativos, como RNN, sequência-para-sequência e redes LSTM, no qual indica a capacidade dessas tecnologias fornecer respostas

precisas aos usuários. A crítica às métricas de avaliação, como BLEU e ROUGE, é destacada, uma vez que sua origem se dá nos sistemas de tradução automática, e já no contexto de *chatbots*, resultam em avaliações com pouca correlação com a avaliação e julgamento realizado por um humano. Isso evidencia a necessidade de desenvolver métricas mais robustas para a avaliação de *chatbots*, um problema de pesquisa que encontra-se em aberto.

Os autores também abordam a questão da acessibilidade para usuários com deficiência auditiva, cognitiva e visual, embora não detalhem explicitamente os requisitos necessários para atender a essa questão. O tema acessibilidade é abordado com foco na usabilidade, isto é, a facilidade dos usuários terem acesso a essas tecnologias, além da capacidade multilíngue dos *chatbots*. Além disso, são levantadas preocupações sobre as limitações semânticas dos *chatbots*, especialmente no reconhecimento de emoções, um fator importante dado os riscos associados às técnicas de aprendizado de máquina não controladas, exemplificados pelo caso do *chatbot* Tay da Microsoft, que se tornou racista, homofóbico e politicamente extremista em suas interações na rede social *Twitter* em 2016.

Ademais, como o estudo compila as principais características e limitações dos *chatbots* entre os anos de 2013 e 2022, desempenha um papel de referência para orientar estratégias de tomada de decisão nesta área. A análise crítica e abrangente contribui para o avanço do conhecimento e a compreensão dos desafios enfrentados no desenvolvimento e uso de *chatbots*, o que contribui para o desenvolvimento deste trabalho.

Os autores Shahin e Ismail (2024) conduziram um estudo utilizando o *ChatGPT* para traduzir a ASL e a LS Árabe. Eles solicitaram ao *ChatGPT* que atuasse como intérprete das linguagens de sinais e realizasse a tradução de ambas as línguas orais para seus respectivos termos em GLOSA. Os resultados dos estudos foram favoráveis, com o *ChatGPT* alcançando uma taxa de sucesso de 100% na tradução da ASL. No entanto, a precisão da tradução para o Árabe foi de apenas 27%. Os autores explicam essa disparidade, sugerindo que o *ChatGPT* pode não ter sido treinado com dados suficientes em GLOSA Árabe, especialmente devido à falta de um dicionário específico para essa LS.

O estudo também destaca as limitações da ferramenta, demonstrando o potencial de seu uso na tradução de línguas orais para a linguagem de sinais escrita, no formato de GLOSA. No entanto, os autores ressaltam que o *ChatGPT* ainda não possuía a capacidade de processar linguagem de sinais através de imagens ou vídeos como entrada. O *link* com este trabalho encontra-se ao também utilizar o *ChatGPT* e o *GEMINI* como uma ferramenta com potencial

para realizar a tradução do Português escrito para GLOSA, de modo a gerar dados para posteriormente treinar uma LLM.

O trabalho de Goldman *et al.* (2025) investiga a capacidade de LLMs em realizar transferência de conhecimento entre idiomas, sem apoio de contexto externo. Para isso, os autores propõem o *benchmark* ECLeKTic, que avalia o desempenho dos modelos em situações em que informações estão disponíveis apenas em um idioma-fonte (por exemplo, inglês), mas não em outros idiomas-alvo (como Português ou tailandês).

A metodologia consiste em selecionar artigos exclusivos de um idioma na *Wikipédia*, gerar perguntas no idioma-fonte e traduzi-las para os demais idiomas. Assim, o modelo precisa responder corretamente em uma língua onde o conteúdo não foi originalmente apresentado, testando sua habilidade de generalizar informações. Foram considerados 12 idiomas, incluindo o Português, e todos os testes seguiram a configuração *zero-shot*, sem ajuste prévio no idioma-alvo.

Os resultados indicam que há queda relevante de desempenho na transferência entre idiomas. O *Gemini 2.0 Pro* apresentou a maior taxa de acerto geral, com 41,3%. Quando se observa a transferência cruzada — isto é, a proporção de respostas corretas no idioma-alvo entre aquelas respondidas corretamente no idioma-fonte — o desempenho foi de 65,0%. Para o Português, a taxa estimada de acerto foi de 67,7%.

Além disso, os experimentos com modelos da série *Qwen 2.5*, em diferentes tamanhos, mostram que o aumento de parâmetros não implica necessariamente melhor transferência de conhecimento, ainda que contribua para um desempenho absoluto mais elevado. Esses resultados reforçam que, embora os LLMs demonstrem capacidade de generalização, ainda há limitações importantes na transferência de conhecimento para idiomas com menor cobertura de dados.

O trabalho é uma referência valiosa nesse contexto, pois investiga abordagens de tradução de bases de dados do inglês para o Português e propõe metodologias de teste em múltiplas variantes: idioma original, idioma quando L2 e variações de tamanho dos modelos, analisando como cada fator impacta a transferência de conhecimento entre idiomas. Em alinhamento a essa perspectiva, esta pesquisa propõe ampliar o escopo de avaliação ao incluir a variante linguística GLOSA, utilizada pela comunidade surda no Brasil. Para isso, serão realizados testes comparativos dos modelos em três contextos: idioma raiz (Inglês), Português como L2 e GLOSA variante da Libras. Além disso, o trabalho prevê o *fine-tuning* dos modelos com dados especificamente convertidos para GLOSA, visando aprimorar a compreensão e geração de texto

nessa variante. Dessa forma, espera-se oferecer evidências mais robustas sobre o potencial dos LLMs para lidar com variações linguísticas pouco representadas e contribuir para soluções mais acessíveis e inclusivas.

O estudo dos autores (Chen *et al.*, 2024) contou com 16 participantes surdos ou com deficiência auditiva para investigar o uso de LLMs na geração de perguntas de quiz com base em respostas emocionais e preferências visuais. As atividades foram elaboradas utilizando formas escritas, vídeos e representações emocionais, integradas ao processo de criação das tarefas. Os resultados indicaram que as estratégias propostas — fundamentadas em dados emocionais e visuais — contribuíram para aprimorar certos aspectos da experiência de aprendizagem, em comparação a métodos tradicionais baseados apenas em transcrições textuais.

Contudo, foram identificados desafios importantes. A diversidade linguística entre os participantes surdos impôs barreiras adicionais ao uso de ferramentas de IA baseadas em texto, sobretudo pela menor proficiência em inglês escrito e pela falta de treinamento específico para utilização dos LLMs. Além disso, os participantes expressaram diferentes níveis de confiança quanto à inclusividade do conteúdo gerado por IA, mencionando preocupações com possíveis vieses linguísticos ou culturais, já que os modelos são treinados principalmente em inglês escrito, bem como com a transparência do uso de dados sensíveis da comunidade surda.

Diante dessas limitações, o estudo recomenda aprofundar pesquisas em duas direções principais: ampliar os conjuntos de dados de treinamento para incluir populações com necessidades linguísticas específicas, como pessoas surdas ou falantes de inglês como segunda língua e aprimorar o design de interação e as estratégias de *prompt engineering* para acomodar perfis diversos de usuários, visando maior acessibilidade e usabilidade.

O trabalho dialoga diretamente com esta pesquisa ao explorar novas abordagens de treinamento voltadas à comunidade surda. Neste estudo, o processo de *fine-tuning* foi conduzido com dados especificamente preparados em GLOSA, buscando instruir os modelos de linguagem a reconhecerem e interpretarem adequadamente essa forma de escrita estruturada. Essa estratégia visa melhorar a capacidade dos LLMs de compreender as particularidades da GLOSA no contexto do Português, contribuindo para reduzir barreiras comunicacionais e ampliando a acessibilidade digital para pessoas surdas.

O artigo (Nunes; Lacerda, 2025) discute o potencial do *ChatGPT* como tecnologia assistiva para o ensino de Português escrito como L2 para alunos surdos, articulando esse uso à perspectiva dos novos letramentos. Por meio de uma análise conceitual, os autores destacam

que o modelo pode atuar como ferramenta de apoio à revisão textual, à explicação gramatical personalizada e à produção de materiais didáticos multimodais, valorizando práticas mais autônomas e inclusivas. No entanto, ressaltam a necessidade de mediação docente constante, dada a diversidade linguística e cultural da comunidade surda, bem como desafios relacionados à dependência excessiva de respostas automatizadas. A reflexão aponta que, embora os LLMs representem uma possibilidade de expansão do letramento em contextos bilíngues, ainda carecem de adequações específicas para atender plenamente as demandas pedagógicas e visuais dos estudantes surdos.

5 AVALIAÇÃO DE CHATBOTS BANCÁRIOS

Para fundamentar esta pesquisa, realizou-se uma avaliação da capacidade de chatbots bancários em compreender e responder a solicitações redigidas em GLOSA. O objetivo central foi identificar as principais dificuldades enfrentadas por esses assistentes virtuais na interação com pessoas surdas. Embora não tenha sido possível verificar diretamente as causas dessas limitações, os resultados observados sugerem que dificuldades na compreensão da GLOSA podem estar associadas à ausência dessa variante nas bases de treinamento utilizadas pelos modelos de linguagem que operam os sistemas avaliados. A metodologia adotada estruturou-se em cinco etapas principais:

1. Simulação de Atividades Bancárias: foram definidos dez cenários baseados em situações reais de atendimento, como consultas de saldo, transferências, pagamentos e solicitação de empréstimos. Esses cenários são apresentados Quadro 5.1

2. Variação nos Textos: para cada cenário, elaboraram-se duas versões de solicitação: uma em linguagem bancária formal e outra em linguagem coloquial, a fim de testar a capacidade dos chatbots de compreender diferentes registros linguísticos.

3. Tradução para GLOSA: os textos foram traduzidos para a variante GLOSA, forma de escrita frequentemente utilizada por pessoas surdas, com base na técnica de glosa sintática. A tradução foi realizada com o suporte do modelo ChatGPT-3.5, e uma amostra representativa foi revisada e validada por um intérprete de Libras para garantir a fidelidade linguística.

4. Interação com Chatbots: as solicitações foram testadas nos chatbots do Banco do Brasil, Bradesco e Itaú, via WhatsApp, por meio de um teste de caixa preta, considerando exclusivamente as respostas fornecidas sem acesso aos processos internos dos sistemas.

5. Avaliação da Compreensão: a análise concentrou-se em verificar se os chatbots eram capazes de compreender corretamente as solicitações e fornecer respostas adequadas. Foram considerados dois critérios: funcionalidade (correção e completude das respostas) e usabilidade (facilidade de uso para o público-alvo).

Quadro 5.1 – Descrição dos Cenários de Interação Bancária

ID	Descrição do Cenário
1	Verificar o saldo da conta
2	Fazer uma transferência
3	Pagar uma fatura ou boleto
4	Verificar as últimas três transações
5	Solicitar o extrato bancário
6	Bloquear ou desbloquear cartão
7	Agendar atendimento presencial
8	Informar-se sobre taxas de juros
9	Relatar perda de cartão roubado ou perdido
10	Solicitar empréstimo pessoal

Fonte: Autora (2025).

Os resultados estão apresentados em Goncalves, Freire e Pereira (2025), e demonstraram que os chatbots possuem desempenho satisfatório na compreensão de solicitações em português formal, mas revelaram limitações significativas quando confrontados com a GLOSA, especialmente na forma coloquial. O chatbot do Itaú obteve o melhor desempenho geral, compreendendo corretamente 90% das solicitações em GLOSA quando redigidas em linguagem bancária formal.

Por outro lado, o Banco do Brasil apresentou uma queda acentuada no desempenho em interações com linguagem coloquial, acertando apenas 30% das solicitações. O Bradesco mostrou inconsistências pontuais, enfrentando dificuldades em cenários específicos, como a solicitação de informações sobre taxas de juros e o bloqueio de cartões.

De modo geral, a taxa de erro variou entre 50% e 80% quando as interações ocorreram em GLOSA, principalmente nas solicitações em linguagem informal. Isso evidencia que pessoas surdas com domínio restrito do português padrão podem encontrar barreiras significativas ao utilizar assistentes virtuais bancários.

A pesquisa revela, portanto, que os chatbots bancários ainda não estão plenamente preparados para atender, de forma inclusiva e eficaz, às necessidades da comunidade surda. Embora haja variações de desempenho entre as instituições avaliadas, observa-se uma tendência geral de maior dificuldade na compreensão da GLOSA, sobretudo quando associada a registros informais de linguagem. Torna-se imprescindível, assim, investir na melhoria da acessibilidade

desses sistemas, a fim de garantir um atendimento bancário verdadeiramente inclusivo, equitativo e eficiente para pessoas surdas.

Foram analisados os erros cometidos pelos *chatbots* tanto em Português formal quanto em *GLOSA*, considerando os dez cenários propostos. No caso da linguagem coloquial, observaram-se falhas no Cenário 08 (solicitação de informações sobre taxa de juros e câmbio), para o Banco do Brasil e Itaú, e no Cenário 10 (solicitação de empréstimo pessoal), para o Banco do Brasil e Bradesco. Já na linguagem bancária formal, os erros ocorreram no Cenário 04 (consulta de saldo e extrato), em ambos Banco do Brasil e Bradesco, além do Cenário 07 (agendamento de atendimento presencial), onde apenas o Bradesco apresentou falha. O destaque positivo foi o *chatbot* do Itaú, que não apresentou erros simultaneamente em Português formal e *GLOSA*, demonstrando maior consistência.

Na análise restrita à *GLOSA*, sintetizada no Quadro 5.2, verifica-se que, em pelo menos oito dos dez cenários avaliados, ocorreu falha em pelo menos uma das plataformas, resultando em taxa de erro de 80% para linguagem coloquial e 50% para linguagem bancária. Assim, a cobertura efetiva de cenários bem atendidos varia entre apenas 20% e 50%, respectivamente.

Quadro 5.2 – Avaliação dos Erros em *GLOSA* conforme os Cenários Propostos

Tipo	**Linguagem Coloquial**	**Linguagem Bancária**
Erros	01, 04, 05, 06, 07, 08, 09, 10	02, 04, 06, 07, 09

Fonte: Autora (2025).

Esses resultados indicam que pessoas surdas enfrentam barreiras significativas ao tentar realizar tarefas bancárias básicas por meio de *chatbots*, sobretudo aquelas com vocabulário restrito ao uso cotidiano e menor familiaridade com terminologias financeiras. O Banco do Brasil destacou-se negativamente, apresentando cerca de 70% de erros na linguagem coloquial, mesmo em solicitações simples, como consultas ou transferências. O Bradesco apresentou desempenho intermediário, mas com inconsistências em cenários críticos, como bloqueio de cartões. Em contrapartida, o Itaú demonstrou maior robustez, alcançando 90% de acertos na linguagem bancária formal e reduzindo falhas mesmo em solicitações coloquiais, embora ainda não atinja plena acessibilidade.

A baixa performance observada está alinhada à falta de representatividade de textos escritos por pessoas surdas nos conjuntos de treinamento dos modelos de linguagem que sustentam esses sistemas. Estudos, como os de Debevc, Kosec e Holzinger (2011), já evidenciam

as dificuldades enfrentadas por pessoas surdas com o Português escrito, o que reforça a necessidade de maior diversidade linguística nos *datasets*. Do ponto de vista teórico, esses achados reforçam tanto a perspectiva do *Design Inclusivo* — que preconiza o desenvolvimento de tecnologias adaptadas a diferentes perfis de usuários — quanto da teoria *Task-Technology Fit* (TTF), a qual sugere que tecnologias só são eficazes quando ajustadas às necessidades reais de seus usuários.

Portanto, observa-se que os *chatbots* avaliados ainda não oferecem suporte adequado às demandas comunicacionais da comunidade surda. Investimentos em modelos de linguagem adaptados à *GLOSA*, seja via *fine-tuning*, seja via pré-processamento linguístico (ex.: mapeamento automático para norma culta), constituem caminhos promissores para ampliar a acessibilidade digital no setor bancário.

6 CONFIGURAÇÃO DOS EXPERIMENTOS

O desenvolvimento deste projeto consistiu na avaliação e treinamento de diferentes LLMs, com o intuito de verificar sua capacidade de interpretar textos na variante GLOSA. Este capítulo corresponde à Etapa 3 da metodologia apresentada na Figura 3.1, que abrange a implementação dos modelos.

Neste contexto, são abordados três dos objetivos específicos definidos na Introdução: (i) verificar a capacidade de modelos de linguagem na interpretação de textos em GLOSA sem ajustes prévios, identificando as principais limitações; (ii) investigar o impacto do *fine-tuning* em diferentes configurações na adaptação dos modelos; e (iii) comparar os resultados obtidos, propondo recomendações para o uso desses modelos em *chatbots* voltados à comunidade surda.

O capítulo está estruturado em cinco seções. A Seção 6.1 apresenta os recursos computacionais utilizados para treinamento e execução dos modelos. A Seção 6.2 descreve as bases de dados empregadas, incluindo as utilizadas para ajuste fino, aprendizado com poucos exemplos e avaliação. Em seguida, a Seção 6.3 detalha a modelagem da solução proposta, abordando desde os processos de tradução e pré-processamento dos dados até a conversão para GLOSA. A Seção 6.4 discute a seleção dos modelos de linguagem utilizados nos experimentos, justificando as escolhas com base em critérios técnicos e de desempenho. Por fim, a Seção 6.5 apresenta o processo de *fine-tuning*, incluindo os parâmetros adotados, as configurações de treinamento e a preparação dos dados de avaliação.

As decisões descritas ao longo deste capítulo foram fundamentais para a análise da viabilidade de adaptação dos LLMs à variante GLOSA, contribuindo diretamente para os objetivos da pesquisa.

6.1 Recursos Computacionais

Neste trabalho, foi utilizada a linguagem de programação *Python*[1] como base para todo o desenvolvimento. O *Python* conta com um ecossistema robusto de bibliotecas e *frameworks* especializados em processamento de linguagem natural e aprendizado de máquina. O *pipeline* do projeto foi estruturado em múltiplas etapas, abrangendo desde o pré-processamento e a tra-

[1] https://www.python.org/

dução dos dados até o treinamento personalizado (*fine-tuning*) dos modelos, além da análise dos resultados e do monitoramento contínuo das métricas de desempenho.

Para a etapa de tradução, a biblioteca *Deep Translator*[2] foi utilizada para converter os textos do inglês para o português de forma automatizada, garantindo maior rapidez e consistência na tradução. O *Gemini 1.5*[3] foi empregado como apoio adicional, auxiliando no pré-processamento dos dados e na tradução para GLOSA, utilizando a abordagem de *few-shot learning*, o que permitiu melhorar a qualidade das traduções e refinar exemplos específicos. Em seguida, o ajuste fino dos modelos de linguagem foi realizado principalmente com as ferramentas da *Hugging Face*[4], *Transformers*[5] e *Unsloth*[6], que possibilitam otimizações de treinamento mais leves e eficientes.

No monitoramento dos experimentos, a plataforma *Wandb*[7] foi o ponto principal para registrar, rastrear e visualizar métricas em tempo real, o que garantiu maior controle sobre o progresso dos treinamentos dos modelos. Além disso, bibliotecas como *Pandas*[8] foram utilizadas para manipulação de dados e organização de *datasets*, enquanto o *Plotly*[9] proporcionou a geração de gráficos interativos que facilitaram a interpretação dos resultados. Para versionamento de código, colaboração e reprodutibilidade, o *GitHub*[10] foi empregado de forma sistemática durante todas as fases do projeto.

Todo o processamento foi executado em um ambiente de hardware robusto, composto por uma GPU NVIDIA GeForce RTX 3090, equipada com 24 GB de memória de vídeo dedicada (VRAM). Essa configuração de alto desempenho foi essencial para viabilizar o treinamento dos LLMs, permitindo uma alocação eficiente dos recursos computacionais via plataforma CUDA — tecnologia de computação paralela da NVIDIA que possibilita o uso intensivo da GPU em tarefas de aprendizado profundo. Com isso, o trabalho foi estruturado para garantir não apenas bons resultados em termos de desempenho, mas também rastreabilidade, segurança dos dados e possibilidade de replicação dos experimentos em pesquisas futuras.

[2] https://pypi.org/project/deep-translator/
[3] [https://deepmind.google/technologies/gemini/
[4] https://huggingface.co/
[5] https://huggingface.co/docs/transformers
[6] https://github.com/unslothai/unsloth
[7] https://wandb.ai/
[8] https://pandas.pydata.org/
[9] https://plotly.com/python/
[10] https://github.com/

6.2 Bases de Dados

Bases de dados são componentes fundamentais para o desenvolvimento de modelos de linguagem de grande porte (LLMs). Para que esses modelos apresentem bom desempenho, é necessário que sejam treinados em conjuntos suficientemente abrangentes e diversificados. Isso se deve ao fato de que os LLMs utilizam técnicas de aprendizado profundo, que requerem grandes volumes de dados para identificar padrões linguísticos e inferir relações entre palavras, frases e conceitos (Zhu, 2024). Além da quantidade, a diversidade da base também contribui para resultados mais generalizáveis: coleções heterogêneas de textos – provenientes de livros, artigos acadêmicos, *websites* e redes sociais – permitem que o modelo aborde uma ampla gama de contextos e perguntas (SCIMAGO INSTITUTIONS RANKINGS, 2021).

Bases de dados bem planejadas auxiliam ainda na mitigação de vieses e inconsistências. Dados enviesados podem gerar respostas discriminatórias ou reforçar estereótipos, enquanto conjuntos balanceados favorecem um desempenho mais ético e alinhado a princípios de responsabilidade (SCIMAGO INSTITUTIONS RANKINGS, 2021). As bases de dados são empregadas em diferentes fases do ciclo de vida dos LLMs:

- **Treinamento**: etapa em que o modelo aprende padrões linguísticos a partir dos dados disponíveis, por exemplo, prevendo a próxima palavra em uma sequência (Zhu, 2024; Sachdeva *et al.*, 2024).
- **Validação e Teste**: conjuntos de dados específicos são reservados para avaliar o desempenho do modelo e ajustar hiperparâmetros conforme necessário (Sachdeva *et al.*, 2024).
- **Ajuste Fino** (*Fine-Tuning*): bases especializadas permitem adaptar o modelo a aplicações ou domínios específicos, tornando-o mais adequado a determinados casos de uso (Sachdeva *et al.*, 2024).

Neste trabalho, diferentes bases de dados são utilizadas em etapas distintas do desenvolvimento. Entre elas, destaca-se a base de dados do INES, descrita em detalhes na subseção 6.2.1, que reúne sinais em Libras e foi adaptada para os experimentos deste estudo, a base de Dados Alpaca, detalhada na subseção 6.2.2, utilizada para a realização do *fine-tuning*, e a base de dados MMLU, apresentada na subseção 6.2.3, que foi utilizada para fazer a avaliação dos experimentos.

6.2.1 *Few-shot Learning*: Dicionário de Dados - INES

A base de dados do INES reúne um total de 5.804 sinais em Libras, organizados para apoiar o aprendizado e a pesquisa na área. Esse material está disponível para consulta por meio do site oficial do *INES*[11].

Para esta pesquisa, a base foi extraída do dicionário online em formato JSON e reorganizada para possibilitar o processamento computacional e a análise com modelos de linguagem. No formato original, o dicionário apresenta diferentes campos informativos; neste trabalho, foram selecionados especificamente os campos **Exemplo (Português)** e **Exemplo LIBRAS/-GLOSA**, estruturados da seguinte forma:

- **Exemplo (Português)** — frase ou definição que indica o significado do termo no contexto do Português.
- **Exemplo LIBRAS/GLOSA** — sequência de palavras em formato simplificado, representando a estrutura da LIBRAS escrita, sem marcas gramaticais do Português, exceto alguns símbolos ocasionais como @ ou ^para indicar flexões.

Essa simplificação mantém a estrutura da língua de sinais, destacando os conceitos-chave sem a necessidade de palavras funcionais do Português. As Figuras 6.1 e 6.2 apresentam exemplos de traduções no dicionário.

Figura 6.1 – Exemplo Dicionário *INES* — Palavra: Computador

Fonte: *INES* (2025).

[11] Base de dados acessível em: http://www.acessibilidadebrasil.org.br/LIBRAS_3/.

Figura 6.2 – Exemplo Dicionário *INES* — Palavra: Educação Física

Fonte: *INES* (2025).

Considerando que o dicionário do INES apresenta, para cada verbete, tanto um **Exemplo em Português** quanto um **Exemplo em LIBRAS/GLOSA**, esta pesquisa utilizou exclusivamente essas frases de exemplo como base para análise.

Foram selecionadas apenas as entradas correspondentes a frases completas, descartando termos isolados ou expressões incompletas. Além disso, optou-se por não incluir frases contendo símbolos adicionais (como ôu), que indicam estruturas mais complexas da LIBRAS e fogem ao escopo desta etapa do estudo.

Entre as frases válidas, selecionaram-se cinco amostras para cada tipo de estrutura frasal — afirmativas, exclamativas e interrogativas — adotando como critério o maior número de palavras, de modo a privilegiar exemplos com mais contexto. Essa escolha visa fornecer entradas mais ricas para a etapa de *few-shot learning*, ainda que o material original seja composto por exemplos relativamente curtos.

As frases selecionadas encontram-se apresentadas nos Quadros 6.1, 6.2 e 6.3.

Quadro 6.1 – Frases Afirmativas em GLOSA/LIBRAS Disponíveis no *INES*.

Id	Português (PT)	GLOSA
4543	Antes eu ganhava pouco, só recebia bolsa de estudo; agora prosperei, ganho bem.	EU ANTES GANHAR POUCO SÓ RECEBER DINHEIRO BOLSA AGORA EU PROSPERAR GANHAR BEM.
4506	No princípio eu não sabia como fazer o dicionário no computador, agora já acostumei.	PRINCÍPIO EU SABER NADA COMO FAZER DICIONÁRIO COMPUTADOR AGORA JÁ ACOSTUMAR.
131	Aqui no Brasil transforma-se a cana em açúcar. Lá na Europa é diferente, transforma-se a beterraba em açúcar.	BRASIL AQUI CANA TRANSFORMAR AÇÚCAR LÁ EUROPA DIFERENTE BETERRABA TRANSFORMAR AÇÚCAR.
3254	Os animais e os bebês precisam tomar leite, é bom para ajudar contra as doenças.	QUALQUER ANIMAL OU BEBÊ PRECISAR TOMAR LEITE BOM AJUDAR CONTRA DOENÇA.
5394	Aqui no Paraná, em julho, sempre há toró. Prefiro viajar para Fortaleza, onde sempre há sol.	PARANÁ JULHO SEMPRE TORÓ EU PREFERIR VIAJAR FORTALEZA LÁ SEMPRE SOL.

Fonte: Autora (2025).

Quadro 6.2 – Frases Exclamativas em GLOSA Disponíveis no *INES*.

ID	Português (PT)	GLOSA
4024	Ontem fui à padaria comprar pão e levei um susto com o aumento do preço!	ONTEM EU IR PADARIA COMPRAR PÃO EU SUSTO PREÇO AUMENTAR!
4865	É melhor retirar um pouco o que você escreveu, eu não gostei!	VOCÊ ESCREVER EU NÃO-GOSTAR MELHOR VOCÊ RETIRAR POUCO!
4135	O carro entrou na São Paulo-Santos e paguei o valor do pedágio com aumento!	SÃO-PAULO SANTOS CARRO ENTRAR PAGAR PEDÁGIO VALOR AUMENTAR!
4679	No Rio de Janeiro há poucos rabinos, em São Paulo há mais!	RABINO RIO TER POUCO SÃO-PAULO TER MAIS!
283	Ontem, eu estava com enxaqueca, tomei comprimido, a cabeça melhorou e aliviou!	ONTEM EU ENXAQUECA TOMAR-COMPRIMIDO CABEÇA MELHOR ALIVIAR!

Fonte: Autora (2025).

Quadro 6.3 – Frases Interrogativas em GLOSA Disponíveis no *INES*.

ID	Português (PT)	GLOSA
305	Você sabe onde fica a loja que vende e aluga apartamentos?	VOCÊ SABER ONDE LOJA CASA APARTAMENTO VENDA ALUGUEL?
4459	Eu tenho biscoitos salgados e doces, qual você prefere?	EU TER BISCOITO SAL DOCE VOCÊ PREFERIR QUAL?
2937	Você vai comprar carro!? Impossível! Onde irá conseguir dinheiro?	VOCÊ VAI COMPRAR CARRO!? IMPOSSÍVEL! ONDE CONSEGUIR DINHEIRO?
3477	Qual você prefere com o pão: manteiga ou queijo?	PÃO VOCÊ PREFERIR MANTEIGA OU QUEIJO QUAL?
4970	O sabor desse suco parece de tangerina ou laranja. Qual é?	SUCO SABOR PARECER TANGERINA OU LARANJA QUAL?

Fonte: Autora (2025).

6.2.2 *Fine-Tuning*: Alpaca

O conjunto de dados *Alpaca* foi desenvolvido por pesquisadores da Universidade de *Stanford* com o objetivo de aprimorar a capacidade dos modelos de linguagem em seguir instruções de forma mais detalhada. Os dados do *Alpaca* seguem o formato *Instruction*, que se refere a um conjunto de dados estruturado especificamente para treinar modelos de linguagem a compreender e executar instruções. Diferentemente de bases de dados tradicionais, que contêm apenas texto genérico, uma base no formato *Instruction* é composta por pares de Instrução e Resposta (Taori *et al.*, 2023).

O propósito desse formato é possibilitar que modelos de linguagem, como *LLaMA*, *GPT*, *GEMINI*, entre outros, sejam treinados para responder a comandos de forma mais útil e alinhada a interações humanas. Segundo Xu *et al.* (2023), esse formato é particularmente relevante para:

- *Chatbots* interativos, como assistentes virtuais.
- Geração de texto controlada, incluindo resumos, explicações e respostas técnicas.
- Modelos de IA especializados, capazes de responder a comandos específicos, como o *ChatGPT*.

A base de dados do *Alpaca* possui aproximadamente 52.000 registros. Os dados utilizados neste estudo foram extraídos do repositório oficial no *GitHub*[12] e traduzidos para o

[12] A base de dados está disponível em: *GitHub Stanford Alpaca*

Português. O 6.4 apresenta exemplos de dados do *Alpaca*. Essa base de dados abrange temas diversos, como Geografia, Programação, Matemática, entre outros. Após o pré-processamento de dados aplicado, o conjunto total de instância utilizado do Alpaca foi de 51.982.

Quadro 6.4 – Exemplos de Dados no Formato Instrução — *Alpaca*.

Instrução	**Resposta**
Explique o que é aprendizado supervisionado em uma frase simples.	Aprendizado supervisionado é um tipo de aprendizado de máquina em que o modelo é treinado com dados rotulados, ou seja, com entradas e saídas conhecidas.
Qual é a capital da França?	A capital da França é Paris.
Escreva um código em *Python* para somar dois números.	`def soma(a, b):` `return a + b`
Descreva a importância da fotossíntese.	A fotossíntese é essencial porque permite que as plantas convertam luz solar em energia, servindo de base para a cadeia alimentar e para a produção de oxigênio.
Liste os primeiros 5 números primos.	Os primeiros 5 números primos são: 2, 3, 5, 7 e 11.
Explique a teoria da relatividade de maneira simples.	A teoria da relatividade afirma que espaço e tempo estão interligados e que a gravidade resulta da curvatura do espaço-tempo ao redor de objetos massivos, como planetas.

Fonte: Autora (2024).

6.2.3 Avaliação: MMLU

O MMLU é um *benchmark* amplamente adotado para avaliação de modelos de linguagem em tarefas de múltipla escolha, abrangendo 57 disciplinas, entre elas matemática, filosofia, direito e medicina. Desenvolvido por Dan Hendrycks e colaboradores em 2020, o MMLU foi projetado para oferecer maior nível de dificuldade em comparação com *benchmarks* anteriores, sendo utilizado em diversos estudos para aferição da compreensão e raciocínio de modelos de linguagem (Hendrycks *et al.*, 2021a; Hendrycks *et al.*, 2021b).

A utilização dessa base de dados possibilita avaliar modelos em configurações de *zero-shot* e *few-shot* (Hendrycks *et al.*, 2021b), o que torna o *benchmark* mais exigente e mais alinhado à forma como o desempenho humano é tradicionalmente avaliado. Além disso, a granularidade e a diversidade das disciplinas abrangidas contribuem para que o MMLU seja especialmente eficaz na identificação de pontos cegos dos modelos.

Para este estudo, a base de dados foi obtida por meio da plataforma *Hugging Face* [13]. Essa base de dados é composta por três diretórios principais — teste, validação e desenvolvimento — além de um diretório adicional de treinamento auxiliar. Os dados estão organizados no formato de perguntas de múltipla escolha, cada uma acompanhada de quatro alternativas de resposta (CAIS, 2023).

O Quadro 6.1 apresenta a organização dos assuntos e a quantidade de instâncias contidas em cada um deles. Neste trabalho, foram consideradas as pastas de desenvolvimento, teste e validação, que foram unificadas, uma vez que, dada a natureza instrucional do problema, não se faz necessária a utilização de um conjunto de validação separado. Após essa fusão, o conjunto final de teste passou a conter 15.573 instâncias, distribuídas conforme os temas indicados no Quadro 6.5.

[13] Disponível em: https://huggingface.co/datasets/cais/mmlu

Quadro 6.5 – Distribuição de Perguntas por Assunto na Base de Dados

Assunto	**Qtd.**	**Assunto**	**Qtd.**
Direito Profissional	1.704	Cenários Morais	995
Assuntos Diversos	869	Psicologia Profissional	681
Psicologia Ensino Médio	605	Macroeconomia Ensino Médio	433
Matemática Fundamental	419	Disputas Morais	384
Pré-História	359	Filosofia	345
Biologia Ensino Médio	342	Nutrição	339
Contabilidade Profissional	313	Medicina Profissional	303
Matemática Ensino Médio	299	Conhecimento Clínico	294
Estudos de Segurança	272	Microeconomia Ensino Médio	264
História Mundial Ensino Médio	263	Física Conceitual	261
Marketing	259	Envelhecimento Humano	246
Estatística Ensino Médio	239	História dos EUA Ensino Médio	226
Química Ensino Médio	225	Sociologia	223
Geografia Ensino Médio	220	Governo e Política Ensino Médio	214
Medicina Universitária	195	Religiões Mundiais	190
Virologia	184	História Europeia Ensino Médio	183
Falácias Lógicas	181	Física Ensino Médio	168
Astronomia	168	Engenharia Elétrica	161
Biologia Universitária	160	Anatomia	149
Sexualidade Humana	143	Lógica Formal	140
Direito Internacional	134	Econometria	126
Aprendizado de Máquina	123	Relações Públicas	122
Jurisprudência	119	Administração	114
Física Universitária	113	Política Externa dos EUA	111
Ética Empresarial	111	Matemática Universitária	111
Ciência da Computação Universitária	111	Segurança da Informação	111
Genética Médica	111	Álgebra Abstrata	111
Fatos Globais	110	Ciência da Computação Ensino Médio	109
Química Universitária	108		

Fonte: Autora (2025).

Vale destacar que os dados apresentados correspondem à base de dados em seu estado bruto. Posteriormente, a base foi submetida a etapas de pré-processamento, definidas pela metodologia adotada, durante as quais algumas instâncias foram descartadas conforme critérios estabelecidos para a análise. Após esse processo, o conjunto final do MMLU utilizado nos experimentos totalizou 15.136 instâncias.

6.3 Modelagem da Solução

A utilização das bases de dados exigiu etapas de processamento distintas, definidas de acordo com a finalidade de cada conjunto. No caso da base MMLU, os dados foram inicialmente traduzidos do inglês para o português e, quando pertinente, convertidos para GLOSA. Além disso, foram identificadas expressões idiomáticas e recursos de linguagem figurativa, de modo a evitar ambiguidades na tradução para GLOSA. Já a base INES, que já contém sinais em Libras, não necessitou de tradução nem de pré-processamento adicional, sendo utilizada diretamente nos experimentos. Esses procedimentos garantem que cada base de dados esteja adequada às etapas específicas do estudo e aos objetivos de cada experimento.

A Figura 6.3 apresenta uma visão unificada do processo adotado nesta pesquisa, integrando as etapas de pré-processamento dos dados e de modelagem/avaliação da solução.

Na etapa 1 (Bases de Dados), diferentes conjuntos de dados foram utilizados, cada um proveniente de fontes específicas. A base INES, contendo sinais em Libras, foi utilizada diretamente nos experimentos, sem necessidade de pré-processamento. A base Alpaca foi empregada na etapa de *fine-tuning* dos modelos, aplicando-se o mesmo pré-processamento utilizado posteriormente na MMLU. Por fim, a base MMLU, composta por textos em inglês, foi traduzida para o português e, quando pertinente, para GLOSA, com ajustes para remover expressões idiomáticas e recursos de linguagem figurativa. Cada base de dados foi utilizada separadamente, de acordo com a etapa e os objetivos do experimento, não havendo fusão entre os conjuntos de dados.

Na etapa 2 (Pré-processamento), códigos de programação, instâncias nulas, *emojis* e expressões matemáticas são preservados em seu formato original, enquanto os demais elementos textuais são traduzidos para o Português e adaptados quanto ao uso de linguagem figurativa. Em seguida, avalia-se se os dados pertencem ao conjunto de treino ou de teste:

- para dados de treino, as perguntas e respostas foram inicialmente traduzidas para o português. Em seguida, realizou-se a remoção de sentidos figurativos presentes nos textos, pois certas expressões idiomáticas ou metáforas podem gerar ambiguidades na representação em GLOSA. Essa etapa envolveu a substituição de expressões figurativas por formas mais literais ou equivalentes claros, garantindo que a tradução para GLOSA fosse precisa e compreensível. O processo foi realizado com o apoio do *Gemini 1.5*, que identificou automaticamente as expressões figurativas e sugeriu substituições adequadas antes da tradução final para GLOSA.

- para dados de teste, as perguntas e respostas são traduzidas para Português, mas apenas as perguntas passam por tradução para GLOSA, pois o objetivo é verificar se os LLMs conseguem interpretar a GLOSA.

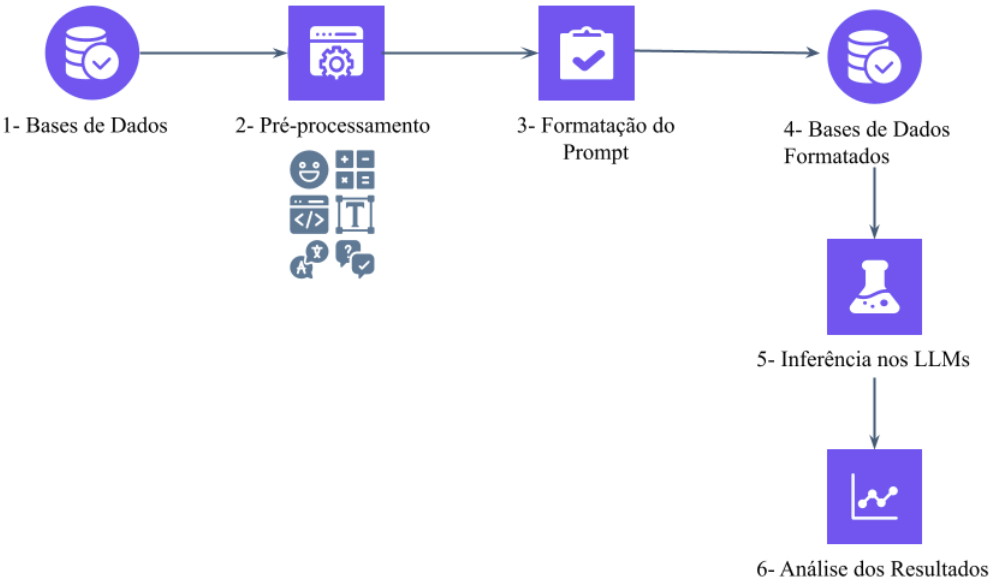
Posteriormente, na etapa 3 (Formatação do *Prompt*), os dados são formatados no estilo *prompt* de instrução (pergunta e resposta), resultando no conjunto final pronto para uso.

Já na etapa 4 (Base de Dados Formatados), os dados são preparados para o *input* nos modelos e cada etapa, seja de treinamento ou de avaliação, contém sua estrutura separada, conforme a estrutura de instrução e de perguntas e respostas.

Na etapa 5 (Inferência nos LLMs), os *prompts* formatados são submetidos aos modelos de linguagem, que produzem as respostas correspondentes.

Por fim, na etapa 6 (Análise dos Resultados), os resultados gerados são compilados e comparados com as respostas originais, permitindo o cálculo das taxas de acerto e erro.

Figura 6.3 – Fluxo de Pré-processamento das Bases de Dados



Fonte: Autora (2025).

6.3.1 Processo de Tradução Inglês/Português

As bases de dados Alpaca e MMLU estão originalmente em *Inglês*, o que torna necessária sua tradução para o *Português* a fim de assegurar acessibilidade e alinhamento com os objetivos deste estudo. Para essa tarefa, foi utilizada a biblioteca *DeepTranslator*[14], uma ferramenta

[14] Disponível em: www.pypi.org/project/deep-translator/

desenvolvida para integrar diferentes APIs de tradução automática, como *Google Translate*, *Microsoft Translator* e *DeepL*. Essa biblioteca foi selecionada em função de sua capacidade de processar diversos pares linguísticos e oferecer traduções satisfatórias para diferentes domínios textuais.

Entretanto, é relevante observar que essas bases de dados não são compostas exclusivamente por textos em linguagem natural. Além de sentenças, incluem códigos de programação, expressões numéricas, *hiperlinks*, imagens e *emojis*. Esses elementos representam desafios adicionais, pois sua tradução inadequada pode comprometer a integridade dos dados.

Para garantir uma tradução fiel, foi necessário identificar previamente as exceções e adotar regras que evitassem sua tradução automática. Dessa forma, trechos contendo códigos de programação, números e símbolos específicos foram isolados e mantidos em sua forma original. Após a tradução do conteúdo restante, esses elementos foram reinseridos na base de dados, preservando sua coerência e evitando distorções na interpretação pelos modelos.

Ademais, mesmo com o uso de uma API de tradução automática, imprecisões podem ocorrer devido a limitações no processamento linguístico, ambiguidades contextuais ou restrições específicas dos tradutores automáticos. Para mitigar esses problemas, foi implementado um processo de verificação automatizado, que consistia no registro de arquivos de *log* a cada instância de tradução processada. Esses *logs* permitiam identificar falhas, como respostas vazias, erros de codificação ou resultados inesperados, possibilitando o reprocessamento seletivo apenas das entradas afetadas. Essa abordagem contribuiu para aumentar a consistência dos dados traduzidos e garantir maior controle de qualidade ao longo do *pipeline*. Quando detectadas, duas estratégias foram consideradas:

1. Reprocessamento da instância para uma segunda tentativa de tradução; ou
2. Exclusão da instância da base de dados, caso a correção não fosse viável.

Um aspecto fundamental desse processo foi buscar a maior completude e coerência possível em cada instância. Para isso, priorizou-se a tradução integral de todos os elementos, incluindo instruções e alternativas de resposta. A fim de preservar a qualidade da base final, definiu-se como critério a exclusão de quaisquer instâncias que apresentassem incongruências não corrigíveis por meio dos processos automatizados de verificação e revisão.

6.3.2 Identificação de Sentido Figurado e Expressões Idiomáticas

O uso de sentido figurado e expressões idiomáticas no contexto da Libras pode gerar ambiguidade e comprometer a qualidade da tradução. Essa questão foi observada durante testes com as ferramentas *Gemini* e *ChatGPT*, utilizadas para verificar a precisão da tradução do *Português* para a GLOSA.

Para aprofundar nesse tipo de análise, este trabalho contou com o apoio de um intérprete especialista em Libras, que confirmou que tais construções linguísticas são fontes recorrentes de inconsistências nas traduções. Isso ocorre porque a Língua de Sinais é interpretativa e altamente dependente do contexto, o que compromete a eficácia da tradução literal de expressões idiomáticas ou de sentido figurado.

Como medida de mitigação de erros, o especialista recomendou a inclusão de uma etapa de pré-processamento, destinada a identificar e adaptar essas expressões, visando uma tradução mais clara e coerente. O Quadro 6.6 apresenta alguns exemplos dessas variações identificadas.

Quadro 6.6 – Análise de Sentidos Figurativos e Possíveis Desvios na Interpretação para Libras – Alpaca

Id	Entrada	Possível Interpretação em Libras
33438	Analise a seguinte citação: "A educação é a arma mais poderosa que você pode usar para mudar o mundo."	O termo "arma" é empregado de forma figurativa para representar o poder da educação. Em Libras, "arma" pode ser interpretado literalmente como objeto bélico, desviando o sentido pretendido de "ferramenta transformadora".
37082	O silêncio foi tão ensurdecedor quanto um trovão.	A palavra "ensurdecedor" expressa um paradoxo na língua portuguesa, pois silêncio implica ausência de som. Em Libras, o termo pode ser traduzido de forma literal, perdendo o efeito de hipérbole e alterando o impacto expressivo da frase.

Fonte: Autora (2025).

Para apoiar essa etapa de identificação, foi desenvolvido um *prompt* específico para o *Gemini-1.5-flash*, instruindo-o a detectar e substituir expressões idiomáticas ou figurativas por alternativas mais apropriadas para a tradução em GLOSA. Esse *prompt* é ilustrado na Figura 6.4.

Figura 6.4 – *Prompt* para Identificação de Expressões Figurativas ou Idiomáticas

```
INSTRUCTION_FIGURATIVE = """Você é um assistente que vai analisar uma base de dados para
                          identificar figuras de linguagem ou expressões idiomáticas.

                          Para cada entrada, você deve:

                          1. Identificar as figuras de linguagem ou expressões idiomáticas na
                          coluna (ORIGINAL).
                          2. Coloque o trecho que represente a figura de linguagem ou expressão na
                          coluna (FIGURATIVO).
                          3. Reescrever o trecho identificado de forma literal (sem a figura de
                          linguagem) em (REESCRITO).
                          4. Substituir o trecho original na frase (ORIGINAL) e coloque em
                          (RESULTADO).
                          5. Retorne somente um json com os campos ID, ORIGINAL, FIGURATIVO,
                          REESCRITO, RESULTADO.
                          """
```

Fonte: Autora(2025).

Após a aplicação do *prompt*, as sentenças foram reformuladas com linguagem direta e literal. Essa abordagem busca tornar as instâncias mais adequadas à tradução para GLOSA, considerando as limitações da Libras na representação de construções ambíguas ou fortemente contextuais.

O Quadro 6.7 apresenta exemplos de resultados obtidos com essa estratégia. A coluna "Original" exibe a instrução inicial; "Expressão Figurativa", o trecho identificado como problemático; "Reescrita", a versão literal proposta; e "Resultado", a sentença final adaptada para uso nos experimentos de tradução.

Quadro 6.7 – Reescrita de Sentenças com Sentido Figurativo para Melhor Compreensão – Alpaca

ID	**Original**	**Expressão Figurativa**	**Reescrita**	**Resultado**
33438	Analise a seguinte citação: "A educação é a arma mais poderosa que você pode usar para mudar o mundo."	"A arma mais poderosa"	A educação é um instrumento muito eficaz que pode ser usado para gerar mudanças no mundo.	Analise a seguinte citação: "A educação é um instrumento muito eficaz que pode ser usado para gerar mudanças no mundo."
37082	O silêncio era ensurdecedor.	ensurdecedor	muito intenso	O silêncio era muito intenso.

Fonte: Autora (2025).

Quadro 6.8 – Exemplos de Frases em Português e Representação em GLOSA – Alpaca

Id	Frase em Português	GLOSA
33438	Analise a seguinte citação: “A educação é um instrumento muito eficaz que pode ser usado para causar transformações no mundo.”	VOCÊ ANALISAR CITAÇÃO EDUCAÇÃO INSTRUMENTO EFICAZ TRANSFORMAÇÃO MUNDO CAUSAR.
37082	O silêncio era muito alto.	SILÊNCIO ALTO FORTE TROVÃO

Fonte: Autora (2025).

Em razão do uso da ferramenta generativa *Gemini* nesta etapa, alguns aspectos éticos, morais e de governança de dados foram observados. Um dos pontos identificados foi a não tradução de determinadas instâncias, resultando no erro *Finish Reason 5*. Consultando a documentação da ferramenta, verificou-se que esse erro se relaciona a possíveis violações das políticas de segurança do *Gemini*. Contudo, não foi possível determinar, até o momento, quais requisitos específicos foram infringidos.

O Quadro 6.9 apresenta exemplos de instâncias para as quais não foi possível identificar ou adaptar sentidos figurativos. Como precaução, essas entradas foram removidas da base de dados. Observa-se que muitos desses casos envolvem termos semanticamente sensíveis, como a palavra ”negro“, o que pode ter contribuído para a detecção automática de risco pela API. A Figura 6.5 representa o erro retornado pela API, uma vez que foi salvo em arquivos de *logs*, para manter o controle do que estava sendo processando corretamente. Para mais informações sobre as diretrizes de segurança da API *Gemini*, recomenda-se consultar a orientação de segurança[15] e as configurações de segurança[16].

[15] Disponível em: https://ai.google.dev/gemini-api/docs/safety-guidance?hl=pt-br.
[16] Disponível em: https://ai.google.dev/gemini-api/docs/safety-settings?hl=pt-br.

Quadro 6.9 – Instâncias Relacionadas a *Finish Reason 5*

Id	Original
2374	Descreva o que é um buraco negro.
9378	O que é um evento Cisne Negro?
10160	Crie uma analogia para ilustrar o conceito de buraco negro.
16885	Crie uma história de terror de três frases que incorpore “corvos negros” e “medos supersticiosos”.
17764	Como você descreveria o conceito de “buraco negro”?
20439	Resuma as características de um buraco negro.
22650	Escreva uma breve descrição das características de um buraco negro.

Fonte: Autora (2025).

Figura 6.5 – Exemplo de Erro Retornado pela API Gemini

```
2024-11-13 12:06:36,899 - ERROR - Ocorreu um erro ao processar a instância 21: ("Invalid operation: The
`response.text` quick accessor requires the response to contain a valid `Part`, but none were returned. The
candidate's [finish_reason](https://ai.google.dev/api/generate-content#finishreason) is 3. The candidate's
safety_ratings are: [category: HARM_CATEGORY_SEXUALLY_EXPLICIT\nprobability: NEGLIGIBLE\n, category:
HARM_CATEGORY_HATE_SPEECH\nprobability: NEGLIGIBLE\n, category: HARM_CATEGORY_HARASSMENT\nprobability:
NEGLIGIBLE\n, category: HARM_CATEGORY_DANGEROUS_CONTENT\nprobability: MEDIUM\n].", [category:
HARM_CATEGORY_SEXUALLY_EXPLICIT
probability: NEGLIGIBLE
, category: HARM_CATEGORY_HATE_SPEECH
probability: NEGLIGIBLE
, category: HARM_CATEGORY_HARASSMENT
probability: NEGLIGIBLE
, category: HARM_CATEGORY_DANGEROUS_CONTENT
probability: MEDIUM
])
```

Fonte: Autora (2025).

6.3.3 Tradução para GLOSA

O processo de tradução para GLOSA teve como objetivo construir uma base de dados em que os textos em português fossem convertidos para a variante GLOSA. No entanto, após diversas pesquisas, não foi identificado um tradutor automático capaz de realizar essa conversão de forma direta.

Diante dessa limitação, adotou-se a técnica de *Few-Shot Learning*, aplicada por meio do modelo *Gemini*, na versão *1.5 Flash*, para realizar as traduções de maneira mais controlada e contextualizada. Também foram conduzidos testes utilizando o *ChatGPT* para essa tarefa; contudo, segundo avaliação do intérprete especialista, os resultados produzidos pelo *Gemini*

apresentaram maior coerência estrutural e melhor desempenho na geração de sentenças adequadas ao formato de GLOSA.

Com base nos testes realizados, foi desenvolvida uma estratégia de *prompt engineering* com o intuito de otimizar as traduções do Português para GLOSA. A Figura 6.6 ilustra o *prompt* elaborado para orientar o modelo na geração dessas traduções. Para apoiar essa abordagem, foram utilizados exemplos no estilo *few-shot*, provenientes dos Quadros 6.1, 6.2 e 6.3, que apresentam sentenças afirmativas, exclamativas e interrogativas respectivamente. O Quadro 6.8, por sua vez, apresenta exemplos práticos das traduções geradas pelo Gemini utilizando esse *prompt*, evidenciando a aplicação da estratégia proposta.

Figura 6.6 – *Prompt* para Tradução em GLOSA

```
INSTRUCTION_GLOSA = """Atue como um tradutor de LIBRAS (Língua Brasileira de Sinais).
                    Textos em português serão enviados e você deve converter para a estrutura de
                    GLOSA, respeitando as seguintes regras:

                    Estrutura de GLOSA mais comum: Objeto + Sujeito + Verbo (OSV) ou Sujeito +
                    Verbo + Objeto (SVO).

                    Exemplo de few-shot learning:
                    ID, PT, GLOSA
                    75, Você recebeu o salário em dinheiro? Acautele-se e esconda-o no sapato.,
                    VOCÊ RECEBER JÁ SÁLARIO DINHEIRO? OLHAR JÁ ESCONDER SAPATO.
                    4543    Antes eu ganhava pouco, só recebia bolsa de estudo; agora prosperei,
                    ganho bem.  EU ANTES GANHAR POUCO SÓ RECEBER DINHEIRO BOLSA AGORA EU
                    PROSPERAR GANHAR BEM.
                    4506    No princípio eu não sabia como fazer o dicionário no computador, agora
                    já acostumei.   PRINCÍPIO EU SABER NADA COMO FAZER DICIONÁRIO COMPUTADOR AGORA
                    JÁ ACOSTUMAR.
                    131 Aqui no Brasil transforma-se a cana em açúcar. Lá na Europa é diferente,
                    transforma-se a beterraba em açúcar.    BRASIL AQUI CANA TRANSFORMAR AÇÚCAR LÁ
                    EUROPA DIFERENTE BETERRABA TRANSFORMAR AÇÚCAR.
                    3254    Os animais e os bebês precisam tomar leite, é bom para ajudar contra
                    as doenças. QUALQUER ANIMAL OU BEBÊ PRECISAR TOMAR LEITE BOM AJUDAR CONTRA
                    DOENÇA.
                    5394    Aqui no Paraná, em julho, sempre há toró. Prefiro viajar para
                    Fortaleza, onde sempre há sol.  PARANÁ JULHO SEMPRE TORÓ EU PREFERIR VIAJAR
                    FORTALEZA LÁ SEMPRE SOL.
                    305 Você sabe onde fica a loja que vende e aluga apartamentos?  VOCÊ SABER
                    ONDE LOJA CASA APARTAMENTO VENDA ALUGUEL ?
                    4459    Eu tenho biscoitos salgados e doces, qual você prefere? EU TER
                    BISCOITO SAL DOCE VOCÊ PREFERIR QUAL?
                    2937    Você vai comprar carro!? Impossível! Onde irá conseguir dinheiro?   VOCÊ
                    VAI COMPRAR CARRO!? IMPOSSÍVEL! ONDE CONSEGUIR DINHEIRO?
                    3477    Qual você prefere com o pão: manteiga ou queijo?    PÃO VOCÊ PREFERIR
                    MANTEIGA OU QUEIJO QUAL?
                    4970    O sabor desse suco parece de tangerina ou laranja. Qual é?  SUCO SABOR
                    PARECER TANGERINA OU LARANJA QUAL?
                    4024    Ontem fui à padaria comprar pão e levei um susto com o aumento do
                    preço!  ONTEM EU IR PADARIA COMPRAR PÃO EU SUSTO PREÇO AUMENTAR.
                    4865    É melhor retirar um pouco o que você escreveu, eu não gostei!   VOCÊ
                    ESCREVER EU NÃO-GOSTAR MELHOR VOCÊ RETIRAR POUCO.
                    4135    O carro entrou na São Paulo-Santos e paguei o valor do pedágio com
                    aumento!    SÃO-PAULO SANTOS CARRO ENTRAR PAGAR PEDÁGIO VALOR AUMENTAR!
                    4679    No Rio de Janeiro há poucos rabinos, em São Paulo há mais!  RABINO RIO
                    TER POUCO SÃO-PAULO TER MAIS!
                    283 Ontem, eu estava com enxaqueca, tomei comprimido, a cabeça melhorou e
                    aliviou!    ONTEM EU ENXAQUECA TOMAR-COMPRIMIDO CABEÇA MELHOR ALIVIAR!

                    Faça o que se pede:
                    1. Traduza a frase reescrita PT para GLOSA e insira o resultado na coluna
                    GLOSA.
                    2. Retorne somente um json com os campos  ID, PT, GLOSA.
                    """
```

Fonte: Autora (2025).

6.3.4 Pré Processamento das Bases de Dados de Avaliação

Como o objetivo é avaliar a capacidade dos modelos de interpretar a GLOSA, adotou-se a seguinte estratégia: as perguntas (*question*) foram traduzidas para GLOSA, enquanto as respostas (*answer*) permaneceram em português. Dessa forma, foi possível verificar se os modelos conseguem compreender corretamente as perguntas em GLOSA e produzir respostas adequadas em português. Para isso, o pré-processamento da base de avaliação incluiu a tradução dos dados

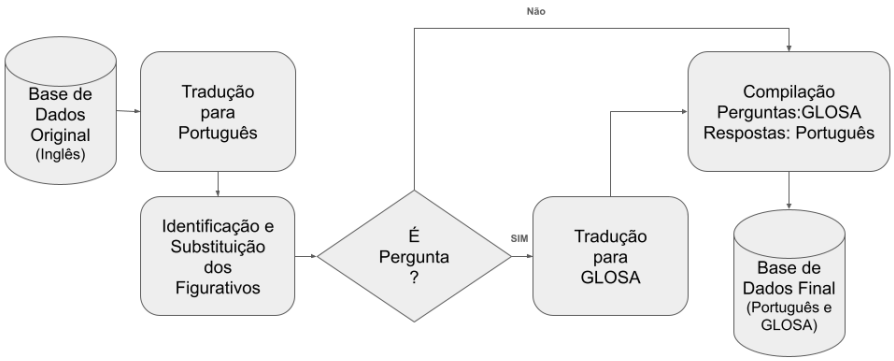
do inglês para o português e, em seguida, a conversão das perguntas do português para GLOSA, conforme descrito nas Subseções 6.3.1, 6.3.2 e 6.3.3.

Essa abordagem visa avaliar a capacidade dos modelos de linguagem em interpretar perguntas formuladas na estrutura linguística da GLOSA, o que é importante para aplicações voltadas à acessibilidade. Um cenário similar ocorre em sistemas como chatbots bancários direcionados a usuários surdos, conforme discutido no Capítulo5. Assim, a estrutura adotada — com perguntas em GLOSA e alternativas em Português — simula um contexto real de interação, no qual um usuário surdo poderia formular sua dúvida em GLOSA e receber respostas estruturadas em Português. O processo de tradução ocorre em duas etapas:

- **Perguntas**: traduzidas do Inglês para o Português e, em seguida, para GLOSA.
- **Alternativas de Resposta**: traduzidas diretamente do Inglês para o Português, sem conversão para GLOSA.

A Figura 6.7 apresenta o fluxo de construção dos dados de avaliação.

Figura 6.7 – Fluxograma do Escopo de Avaliação



Fonte: Autora (2025).

A formatação do *prompt* da instrução que será passada para o LLM é fundamental para garantir que ao avaliar o LLM, seja possível obter respostas coerentes e consistentes. Para isso, após a tradução e estruturação das perguntas e alternativas, os dados são organizados para servir como *prompts* de instrução. Cada *prompt* é composto por um contexto, uma pergunta e as alternativas de resposta, sendo padronizado para todas as instâncias. A forma como os

dados são apresentados aos modelos tem grande impacto na precisão dos resultados obtidos, pois influencia diretamente como os modelos interpretam e geram respostas.

A estrutura do *prompt* foi definida de forma a assegurar que todos os modelos recebam os mesmos dados, eliminando possíveis vieses na apresentação das informações. A Figura 6.8 apresenta o formato do *prompt* utilizado em cada instância. A Figura 6.9 mostra um exemplo de *input* após a formatação. A resposta correta para essa pergunta segundo a base de dados de respostas, é a letra (A).

Figura 6.8 – *Prompt* de Instrução *Avaliação*

```
prompt = """Você é um assistente altamente capacitado. Leia a pergunta com atenção e escolha a
          resposta correta entre as opções (A), (B), (C) ou (D).
          ### Pergunta:
          ### Opções Disponíveis:
          ### Resposta:
        """
```

Fonte: Autora (2025).

Figura 6.9 – *Prompt* de Instrução *Avaliação* Completo - Exemplo

```
prompt = """Você é um assistente altamente capacitado. Leia a pergunta com atenção e escolha a
          resposta correta entre as opções (A), (B), (C) ou (D).
          ### Pergunta: HOMEM JOVEM ADULTO PESO NORMAL COMPOSIÇÃO CORPORAL QUAL?
          ### Opções Disponíveis:
          (A) Gordura = 17%, Massa livre de gordura = 83%, Fluido intracelular = 40%, Fluido
              extracelular = 20%
          (B) Gordura = 83%, Massa livre de gordura = 40%, Fluido intracelular = 17%, Fluido
              extracelular = 20%
          (C) Gordura = 20%, Massa Livre de Gordura = 17%, Fluido Intracelular = 40%, Fluido
              Extracelular = 83%
          (D) Gordura = 40%, Massa Livre de Gordura = 20%, Fluido Intracelular = 17%, Fluido
              Extracelular = 83%
          ### Resposta:
        """
```

Fonte: Autora (2025).

6.4 Seleção de Modelos para *Fine-Tuning*

A seleção dos modelos de linguagem utilizados neste trabalho baseou-se em critérios técnicos e pragmáticos. Optou-se por modelos de código aberto, compatíveis com o idioma Português e viáveis para ajustes por *fine-tuning* em uma máquina local, sem custos adicionais. A principal motivação foi garantir que os modelos pudessem ser adaptados à GLOSA.

Todos os modelos selecionados estão disponíveis na plataforma *Hugging Face*, o que viabiliza sua integração em pipelines de PLN de forma simples e reprodutível (Wolf *et al.*, 2020). Além disso, são compatíveis com o *framework Unsloth*, uma ferramenta que permite realizar inferência e *fine-tuning* de LLMs. Isso torna possível seu uso até mesmo em ambientes com recursos computacionais limitados, como notebooks ou GPUs de médio porte (Daniel; Michael Han; team, 2023).

Outro critério decisivo foi a capacidade multilíngue dos modelos, com suporte ao Português, considerando que as tarefas abordadas envolvem diretamente esse idioma. Além disso, o limite de tamanho dos modelos selecionados foi diretamente influenciado pelas restrições dos recursos computacionais disponíveis para os experimentos, priorizando arquiteturas que pudessem ser executadas localmente com eficiência. Os principais critérios de escolha foram:

- Disponibilidade pública e acesso facilitado via *Hugging Face*;
- Compatibilidade com o *Unsloth*, garantindo eficiência tanto na inferência quanto no *fine-tuning*;
- Capacidade multilíngue, com suporte ao Português;
- Viabilidade de execução em ambiente local, considerando limitações de GPU e memória.

O Quadro 6.10 apresenta uma visão geral das principais características dos modelos de linguagem selecionados para este trabalho. Ressalta-se que o "tamanho" mencionado refere-se ao número de parâmetros de cada modelo, o que impacta diretamente sua capacidade de generalização, necessidade de recursos computacionais e tempo de treinamento. As informações listadas foram obtidas nos documentos oficiais e repositórios públicos de cada modelo.

Complementarmente, o Quadro 6.11 destaca os principais pontos fortes e limitações identificados durante a análise desses modelos, com base tanto na documentação técnica quanto na experiência prática observada durante os experimentos.

Quadro 6.10 – Visão geral dos Modelos Selecionados

Família	Versão	Tam.	Idiomas	Dados de Treinamento	Melhor Desempenho
LLaMA	Llama-3.1-8B-unsloth-bnb-4bit	8B	Multilíngue	Dados abertos supervisionados	Inferência geral, reescrita.
LLaMA	Llama-3.2-3B-Instruct-unsloth-bnb-4bit	3B	Multilíngue	Dados abertos supervisionados	Tradução, janelas longas.
Phi	phi-4-unsloth-bnb-4bit	4B	Multilíngue	Dados didáticos e sintéticos curados	Raciocínio lógico, geração de código
Qwen	Qwen2.5-14B-Instruct-unsloth-bnb-4bit	14B	Multilíngue	Corpora multilíngue	QA, geração de texto, tradução
Qwen	Qwen2.5-7B-Instruct-bnb-4bit	7B	Multilíngue	Corpora multilíngue	Tarefas multilíngues com menos recursos
Zephyr	zephyr-sft-bnb-4bit	7B	Multilíngue	Preferência sintética via dDPO	Diálogo, alinhamento de instruções

Fonte: Autora (2025).

Quadro 6.11 – Pontos Fortes e Limitações dos Modelos Selecionados

Versão	Pontos Fortes	Limitações
LLaMA 3.1-8B-unsloth-bnb-4bit	Equilíbrio entre tamanho e qualidade, bom para PT, janelas longas	Maior demanda de VRAM, restrição comercial para usos sensíveis
LLaMA 3.2-3B-Instruct-unsloth-bnb-4bit	Janelas de contexto extensas, multimodal, base sólida para PT	Tamanho reduzido limita raciocínio mais complexo
Phi-4-unsloth-bnb-4bit	Compacto, bom em raciocínio lógico, baixo custo computacional	Contexto curto, menor robustez para multilíngue e PT
Qwen2.5-14B-Instruct-unsloth-bnb-4bit	Robusto para multilíngue, adapta-se bem em QA e geração, open-source	Alta demanda de VRAM, requer ajuste fino para PT avançado
Qwen2.5-7B-Instruct-bnb-4bit	Flexível para hardware modesto, bom suporte multilíngue	Desempenho inferior ao 14B em tarefas complexas
Zephyr-sft-bnb-4bit	Leve, eficiente, alinhamento forte via dDPO, rápido para chat	Limitado em multilíngue, raciocínio técnico restrito

Fonte: Autora (2025).

A seguir, é detalhado cada um dos modelos selecionados — *LLaMA*, *Phi*, *Qwen* e *Zephyr* — destacando suas características principais e os motivos de sua inclusão neste estudo.

6.4.1 LLaMa

A família de modelos LLaMA é desenvolvida pela *Meta AI* desde fevereiro de 2023, consiste em modelos de base que variam de 3 a 65 bilhões de parâmetros, treinados com trilhões de *tokens* provenientes exclusivamente de fontes públicas (Touvron *et al.*, 2023). Um aspecto importante do LLaMA é sua eficiência em termos de inferência o que possibilita a execução em uma única GPU, assim, viabiliza seu uso em ambientes com recursos computacionais mais limitados. Desde o lançamento, os modelos foram disponibilizados abertamente à comunidade de pesquisa, o que fomenta o avanço colaborativo na área de modelos de linguagem.

Em continuidade à série, o *LLaMA 3* representa uma evolução em relação às versões anteriores, e com variantes recentes, como as versões 3.1 e 3.2. Essas variantes foram selecionadas para experimentação neste trabalho por apresentarem desempenho aprimorado, suporte multilíngue consistente e facilidade de integração em *pipelines* de inferência e *fine-tuning*. Um ponto de destaque é a maior capacidade de contexto, com janelas de até 128 mil *tokens*, o que

viabiliza o processamento de textos extensos ou diálogos prolongados de forma mais coesa (AI, 2024). Além disso, o LLaMA 3.2 introduz variantes multimodais, permitindo entradas de texto e imagem, ampliando o leque de aplicações para tarefas mais complexas.

Entre os pontos positivos, destacam-se o desempenho competitivo em diversas tarefas, superando inclusive modelos de maior porte em vários *benchmarks* de geração de texto, compreensão e raciocínio. A família também demonstra flexibilidade para adaptação a diferentes contextos, que oferece versões otimizadas para dispositivos com menor capacidade de processamento, no qual amplia sua aplicabilidade em cenários com restrições de *hardware*. Por outro lado, algumas limitações devem ser observadas: variantes de maior porte podem exigir infraestrutura computacional mais robusta, como GPUs de alto desempenho, para operar em sua capacidade total. A capacidade multimodal, embora já presente, ainda se encontra em estágio inicial e pode demandar melhorias conforme a complexidade das tarefas.

6.4.2 Phi

A família *Phi* é desenvolvida pela *Microsoft Research* e foi concebida com o objetivo de oferecer modelos de linguagem compactos, eficientes e de qualidade satisfatória, mesmo em ambientes com recursos computacionais limitados. A abordagem central da série segue a filosofia "*Small yet Capable*", cujo foco é priorizar modelos menores, porém treinados com dados criteriosamente selecionados e instruções de caráter didático, buscando maximizar a utilidade do conhecimento aprendido (Chen *et al.*, 2023).

O *Phi-1* foi o primeiro modelo da série, possui 1,3 bilhões de parâmetros e foi projetado com foco na geração de código. Diferentemente de modelos generalistas de maior porte, o *Phi-1* foi treinado em dados descritos como de "qualidade de livro didático" (Chen *et al.*, 2023). Esse *corpus* inclui aproximadamente 6 bilhões de *tokens* extraídos da *web* e cerca de 1 bilhão de *tokens* provenientes de exercícios sintéticos, elaborados com apoio do *GPT-3.5*.

Apesar do tamanho reduzido, o *Phi-1* apresentou características de raciocínio e compreensão de tarefas geralmente observadas em modelos mais robustos. Sua arquitetura enxuta também favorece maior eficiência, o que demanda menos *tokens* de entrada e menor utilização de recursos computacionais durante o treinamento.

Dando continuidade à proposta, a *Microsoft* lançou o *Phi-4*, com 14 bilhões de parâmetros. Em alinhamento com a filosofia da família *Phi* o modelo mantém o foco na qualidade

dos dados e incorpora dados sintéticos durante o pré e o *midtraining* (Abdin *et al.*, 2024). Essa síntese é gerada por técnicas como *multi-agent prompting*, auto-revisão e revisão de instruções, complementada por dados orgânicos filtrados (como *web*, livros licenciados e repositórios de código). O treinamento inclui uma transição de currículos e refinamentos pós-treinamento, utilizando rejeição de amostras indesejadas e uma abordagem de *Direct Preference Optimization* (*DPO*) (Abdin *et al.*, 2024).

Em termos de aplicação prática, a família *Phi* apresenta boa eficiência com demanda reduzida de *VRAM*, o que viabiliza sua utilização em *frameworks* como o *Unsloth*, mesmo em ambientes com restrições de hardware. Além disso, demonstra desempenho competitivo em tarefas de raciocínio lógico, interpretação de instruções e geração de texto coerente, mantendo resultados consistentes mesmo em dispositivos com menor capacidade.

Outro aspecto relevante é sua disponibilidade pública na plataforma *Hugging Face*, o que facilita a integração, inferência e *fine-tuning* em fluxos de trabalho de processamento de linguagem natural. Por outro lado sua capacidade de memória de contexto também é mais restrita em comparação a modelos maiores, o que limita o processamento de textos extensos em uma única janela. Ademais, a dependência de dados sintéticos durante o treinamento pode resultar em lacunas de conhecimento em domínios muito específicos, tornando necessários ajustes complementares para tarefas mais especializadas (Abdin *et al.*, 2024).

6.4.3 Qwen

A família *Qwen*, desenvolvida pela Alibaba Cloud, constitui uma série de modelos de linguagem de grande porte com suporte multilíngue e capacidades multimodais (Team, 2023). Desde seu lançamento, os modelos Qwen vêm sendo reconhecidos por sua capacidade de compreensão e geração de texto em diversos idiomas, incluindo o Português.

A linha Qwen compreende modelos de diferentes tamanhos e varia de versões mais compactas (7B e 14B parâmetros) a arquiteturas mais robustas, com até 32B parâmetros. Os modelos mais recentes incorporam funcionalidades multimodais, sendo capazes de processar não apenas texto, mas também entradas de áudio, vídeo e elementos de visão computacional.

Uma característica relevante da família Qwen é a combinação de treinamento em grandes corpora multilíngues com técnicas contemporâneas de ajuste fino, como *Supervised Fine-Tuning* e *Reinforcement Learning from Human Feedback* (*RLHF*). Essa abordagem busca pro-

mover maior alinhamento às instruções humanas o que favorece a geração de respostas consistentes (Team, 2024; Team, 2025). As versões mais recentes como o Qwen 2.5 e o Qwen 3, incluem melhorias nos processos de pré-treinamento, estratégias de filtragem de conteúdo indesejado e heurísticas voltadas à mitigação de vieses e alucinações, seguindo diretrizes para o uso responsável de LLMs de código aberto, licenciados sob a Apache-2.0.

Entre os aspectos positivos da família Qwen, destaca-se o suporte multilíngue que abrange idiomas de alta e média frequência, como o Português, e o desempenho satisfatório em tarefas de *question answering*, geração de texto e raciocínio lógico. A disponibilidade pública dos modelos no *Hugging Face* e a compatibilidade com *frameworks* como o *Unsloth* também permitem a realização de inferências e *fine-tuning* em diversos ambientes computacionais. Por outro lado, apesar do suporte multilíngue, o desempenho mais consolidado ainda ocorre em Inglês e Mandarim, sendo recomendável realizar ajustes adicionais para aprimorar resultados em Português.

6.4.4 Zephyr

O *Zephyr-7B* é um modelo de linguagem de código aberto desenvolvido com foco em diálogos interativos e seguimento de instruções de forma natural e coerente. O *Zephyr* integra uma série de modelos ajustados para atuar como assistentes virtuais úteis e responsivos (Tunstall *et al.*, 2023).

Um dos diferenciais do *Zephyr* está no processo de alinhamento baseado na técnica *Distilled Direct Preference Optimization* (*dDPO*), conforme descrito por Rafailov *et al.* (2023). Nessa abordagem, os dados são gerados de preferência por um modelo "professor" para treinar o modelo "aluno", assim, dispensando a necessidade de anotações humanas extensivas. Essa estratégia contribui para reduzir custos e tempo de treinamento, ao mesmo tempo em que busca aprimorar a capacidade do modelo de interpretar instruções e responder de acordo com a intenção do usuário.

Apesar de contar com apenas 7 bilhões de parâmetros o *Zephyr-7B* apresenta desempenho competitivo em avaliações de conversação, como no *MT-Bench*, que ele supera em alguns casos modelos de maior porte como o *LLaMA2-Chat-70B*. Além disso, seu treinamento com *dDPO* é relativamente mais ágil e reprodutível, pois não depende de pipelines de *RLHF* complexos ou processos de amostragem onerosos (Latenode, 2024).

O modelo *Zephyr* destaca-se também por sua eficiência em ambientes com recursos computacionais limitados, sendo indicado para aplicações que operam sob restrições de *VRAM* ou infraestrutura. Outro ponto positivo é seu caráter totalmente aberto: código, pesos e tutoriais estão amplamente disponíveis, o que favorece a reprodutibilidade e a customização em diferentes contextos. Adicionalmente, o *Zephyr* demonstra alinhamento consistente a instruções, oferecendo respostas adequadas em tarefas conversacionais e interações baseadas em diálogo.

Por outro lado, por não ter sido projetado originalmente como um modelo multilíngue, seu desempenho é mais sólido em Inglês, podendo demandar ajustes ou *fine-tuning* adicionais para aplicações em Português. Além disso, seu foco em tarefas de *chat* pode limitar sua eficácia em domínios que exijam raciocínio técnico mais detalhado ou geração de código especializada.

6.5 Ajuste Fino (*Fine-Tuning*)

O *fine-tuning*, ou ajuste fino, é uma técnica de aprendizado de máquina utilizada para especializar modelos de linguagem pré-treinados em tarefas ou domínios específicos. Nesse processo, o modelo, inicialmente treinado em grandes *corporas* de dados gerais ou genéricos, é posteriormente adaptado a um contexto mais restrito, incorporando características linguísticas, semânticas e estilísticas que não foram totalmente capturadas no treinamento original (Radford *et al.*, 2018b).

No contexto deste trabalho, o *fine-tuning* desempenha um papel fundamental para tornar o modelo capaz de interpretar textos em Português segundo a variante GLOSA. Embora os modelos de linguagem pré-treinados, como *Qwen*, *Phi*, *Llama* ou *Zephyr*, já tragam conhecimento linguístico amplo adquirido a partir de grandes *corporas* de livros, sites e textos diversos, eles não estão preparados para lidar com as especificidades da GLOSA. Assim, o ajuste fino permite refinar os parâmetros do modelo para que ele assimile padrões sintáticos, regras gramaticais simplificadas e nuances próprias dessa variante.

Além de possibilitar uma *customização* linguística mais precisa, o *fine-tuning* é também uma abordagem eficiente do ponto de vista computacional. Treinar modelos de linguagem do zero é uma tarefa extremamente custosa, exigindo enormes quantidades de dados, processamento intensivo e longos períodos de treinamento. Com o ajuste fino, aproveita-se o aprendizado já incorporado na etapa de *pré-treinamento*, economizando recursos e tempo, mas ga-

rantindo que o modelo se torne altamente especializado para a tarefa proposta (Radford *et al.*, 2018b).

Para viabilizar essa adaptação, foi empregada uma combinação de ferramentas robustas do ecossistema *Python*. Bibliotecas como *Hugging Face* e *Unsloth* foram fundamentais para estruturar o *pipeline* de treinamento, possibilitando otimizações de memória e tempo de processamento. O monitoramento do processo foi realizado por meio do *Wandb*, garantindo rastreabilidade das métricas em tempo real e facilitando ajustes de *hiperparâmetros*. Todo o processo foi executado em uma GPU *NVIDIA GeForce RTX 3090*, com 24 GB de memória dedicada, fator determinante para lidar com modelos de grande porte e etapas de retro propagação de gradientes complexas.

Por fim, para garantir a qualidade e a confiabilidade dos resultados, o trabalho foi dividido em versões de configuração distintas, cada uma explorando diferentes combinações de parâmetros de treinamento, tamanho de *batch* e número de épocas. Essa abordagem permitiu avaliar o equilíbrio entre eficiência computacional e robustez linguística de tal modo que o modelo fosse não apenas funcional, mas também replicável e adaptável para futuras extensões ou aplicações práticas.

6.5.1 Principais Parâmetros

O treinamento de LLMs envolve uma série de parâmetros ajustáveis que impactam diretamente a eficiência computacional, a qualidade do aprendizado e a capacidade de generalização. A configuração adequada desses parâmetros é fundamental para alcançar um equilíbrio entre tempo de processamento, uso de memória e qualidade das respostas geradas (Chawre, 2023). Nesta seção, são apresentados os principais parâmetros utilizados no *fine-tuning* e suas justificativas, bem como a comparação entre as duas abordagens de configuração aplicadas neste trabalho.

- **Número de Épocas (*num_train_epochs*):** define quantas vezes o modelo percorre todo o conjunto de treinamento. Um valor maior tende a melhorar o aprendizado, mas pode levar ao *overfitting* caso exceda o ponto ótimo.

- **Tamanho Máximo de Sequência (*max_seq_length*):** estabelece o número máximo de tokens processados por entrada. Sequências mais longas capturam mais contexto, mas aumentam significativamente o consumo de memória.

- **Taxa de Aprendizado (*learning_rate*):** controla o tamanho dos ajustes nos pesos do modelo durante cada atualização. Taxas mais altas aceleram o treinamento, mas podem gerar oscilações; taxas muito baixas tornam o aprendizado mais estável, porém mais lento.

- **Passos de Aquecimento (*warmup_steps*):** definem uma fase inicial com taxa de aprendizado gradual para evitar variações bruscas no início do treinamento.

- **Tamanho de Lote por Dispositivo (*per_device_train_batch_size*):** indica quantas amostras são processadas simultaneamente por GPU. Valores maiores reduzem o tempo de treinamento, mas exigem mais memória.

- **Acumulação de Gradiente (*gradient_accumulation_steps*):** permite acumular gradientes de vários *batches* antes de atualizar os pesos, viabilizando o uso de *batch sizes* efetivos maiores em GPUs com menor capacidade de memória.

- **Decay de Peso (*weight_decay*):** adiciona uma forma de regularização para evitar *overfitting*, penalizando pesos excessivamente grandes.

- **Frequência e Limite de Salvamento (*save_steps* e *save_total_limit*):** esses parâmetros definem a cada quantos passos de treinamento os *checkpoints* do modelo são salvos (*save_steps*) e o número máximo de *checkpoints* que podem ser mantidos (*save_total_limit*). Isso permite recuperar o modelo em caso de falhas, sem ocupar espaço desnecessário em disco, equilibrando segurança e eficiência de I/O.

6.5.2 Configurações de Treinamento

Para avaliar o impacto das diferentes combinações de parâmetros, foram adotadas duas estratégias de configuração denominadas Versão 1 (*V1*) e Versão 2 (*V2*), cada uma com objetivos distintos. O Quadro 6.12 sintetiza os parâmetros escolhidos para cada versão.

A configuração adotada baseou-se em códigos de exemplo previamente disponíveis para os modelos utilizados, bem como nas recomendações presentes em guias técnicos voltados ao

ajuste fino de modelos de linguagem. Destacam-se, nesse contexto, os materiais disponibilizados pela comunidade Hugging Face[17] e pela biblioteca Unsloth[18], que oferecem orientações consolidadas sobre práticas de treinamento eficientes, com ênfase na redução do custo computacional e na obtenção de resultados consistentes em ciclos curtos de experimentação.

- **Versão 1 (*V1*):** prioriza rapidez e uso eficiente de recursos computacionais, utilizando menos épocas, maior taxa de aprendizado e sequência de contexto moderada. Ideal para iterações rápidas e prototipagem.
- **Versão 2 (*V2*):** foca em robustez e qualidade de generalização, empregando mais épocas, taxa de aprendizado mais conservadora e sequência de contexto mais longa. Requer maior poder de processamento e tende a gerar modelos mais estáveis.

Quadro 6.12 – Parâmetros das Configurações de Treinamento (*V1* e *V2*).

Parâmetro	***V1* (Rápida)**	***V2* (Robusta)**
num_train_epochs	2	5
max_seq_length	4096	8192
learning_rate	2e-4	1e-4
warmup_steps	50	500
per_device_train_batch_size	2	4
gradient_accumulation_steps	8	8
weight_decay	0.01	0.02
save_steps	200	500
save_total_limit	3	5

Fonte: Autora (2025).

A comparação entre as duas versões demonstra como as escolhas de parâmetros impactam diretamente o processo de *fine-tuning*. A configuração com menor tempo de processamento é recomendada para ciclos de experimentação e ajustes preliminares, enquanto a versão mais robusta é indicada para ambientes de produção ou quando se busca maior capacidade de generalização do modelo.

[17] https://huggingface.co/docs/transformers/training
[18] https://github.com/unslothai/unsloth

7 ANÁLISE DOS RESULTADOS

Este capítulo apresenta os principais resultados obtidos na avaliação dos LLMs em diferentes conjuntos de dados. O objetivo principal é analisar o desempenho dos modelos em bases de dados originais e em suas variantes, considerando os processos de *fine-tuning* aplicados a cada configuração. Foram utilizados quatro conjuntos de dados: *Inglês* (original), **Português**, **GLOSA e Português** (misto) e **GLOSA** (puro), todos no contexto de *Question Answering*.

A avaliação foi estruturada para verificar se o *fine-tuning* contribui para aprimorar a capacidade dos modelos de compreender e gerar respostas consistentes no domínio linguístico restrito da GLOSA. Para isso, definiu-se uma estratégia experimental composta por múltiplos cenários de teste, contemplando não apenas o desempenho geral nos idiomas de origem, mas também a adaptação dos modelos à variações linguísticas de complexidade crescente.

Cada modelo selecionado — *LLaMA*, *Qwen*, *Phi* e *Zephyr* — foi avaliado em três versões distintas:

- **Versão base**: modelo original, pré-treinado em textos multilíngues, sem aplicação de *fine-tuning*.
- **V1**: modelo com treinamento adicional (*fine-tuning*) priorizando agilidade e eficiência no uso de recursos computacionais.
- **V2**: modelo com treinamento adicional (*fine-tuning*) voltado à robustez e qualidade das respostas geradas.

Essa abordagem permitiu analisar o impacto real do *fine-tuning* na adaptação dos modelos a contextos linguísticos específicos. A diversidade dos conjuntos de dados foi essencial para verificar como o desempenho varia de acordo com o grau de familiaridade do modelo com o idioma e sua estrutura sintática. Foram consideradas quatro variações principais:

- **Inglês**: dados em Inglês, idioma predominante no treinamento original dos modelos e fonte primária dos dados, ou seja, a base de dados no seu estado original.
- **Português**: dados em Português padrão, para avaliação da competência multilíngue.
- **GLOSA e Português**: tradução controlada, na qual as perguntas seguem as regras da GLOSA e as respostas são apresentadas em Português, simulando um cenário de aplicação real.

- **GLOSA**: perguntas e alternativas integralmente reescritas em GLOSA, representando o nível máximo de complexidade interpretativa.

A combinação de diferentes modelos, versões e conjuntos de dados possibilitou verificar em que medida os LLMs são capazes de generalizar padrões sintáticos e semânticos não convencionais, identificar limitações na interpretação e quantificar os ganhos obtidos com o *fine-tuning*. O *fine-tuning* dos modelos foi realizado exclusivamente com os dados do conjunto Alpaca, enquanto a avaliação de desempenho foi conduzida com base nos dados do *benchmark* MMLU. É importante destacar que, de acordo com a documentação dos modelos, eles não tiveram acesso aos conjuntos de treinamento durante seu pré-treinamento, ou seja, embora o *benchmark* disponha de uma partição de treinamento, os modelos não foram expostos previamente a esses dados, garantindo que a avaliação seja realizada em exemplos inéditos e evitando contaminação. Esse mapeamento é fundamental para orientar futuras melhorias nos processos de treinamento e aplicação desses modelos em contextos linguísticos específicos, como o uso da GLOSA.

7.1 Visão Geral da Avaliação

O Quadro 7.1 apresenta a síntese dos cenários de teste avaliados, acompanhados dos cenários formulados e suas respectivas justificativas. Esses cenários foram definidos para explorar diferentes dimensões da adaptação dos modelos a variações linguísticas e estratégias de *fine-tuning*.

Quadro 7.1 – Cenários avaliados e resultados esperados *fine-tuning*.

Cenário	**Resultado Esperado**
Modelo: Original Dados: Inglês	Espera-se desempenho máximo, pois os modelos foram majoritariamente pré-treinados neste idioma. Este cenário serve como *baseline* para medir o impacto do idioma e do *fine-tuning*, além de validar a robustez do *prompt*.
Modelo: Original Dados: Português	Espera-se desempenho inferior ao inglês, evidenciando limitações na competência multilíngue do modelo sem ajustes específicos. Permite identificar desafios sintáticos e semânticos no Português.
Modelo: Original Dados: Perguntas em GLOSA Respostas em Português	Espera-se desempenho moderado a baixo, devido à dificuldade de interpretar GLOSA sem *fine-tuning*, simulando um cenário real de tradução controlada.
Modelo: Original Dados: Perguntas em GLOSA Respostas em GLOSA	Espera-se desempenho muito baixo, pois GLOSA é ausente no pré-treinamento do modelo. Este cenário evidencia a necessidade de *fine-tuning* para variantes linguísticas controladas.
Modelo: Ajustado V1 Dados: Perguntas em GLOSA Respostas em Português	Espera-se melhora parcial no desempenho, já que a estratégia V1 de *fine-tuning* prioriza eficiência de tempo e recursos, ainda que com possível perda de robustez.
Modelo: Ajustado V2 Dados: Perguntas em GLOSA Respostas em Português	Espera-se melhora superior no desempenho, pois a estratégia V2 busca maior robustez e qualidade, ainda que com maior custo computacional. A expectativa é que o ganho de desempenho compensará o custo adicional.

Este desenho experimental busca uma comparação sistemática e reprodutível, permitindo identificar limitações e oportunidades para aprimorar a adaptação dos modelos a variantes linguísticas específicas. As seções subsequentes detalham os resultados para cada cenário.

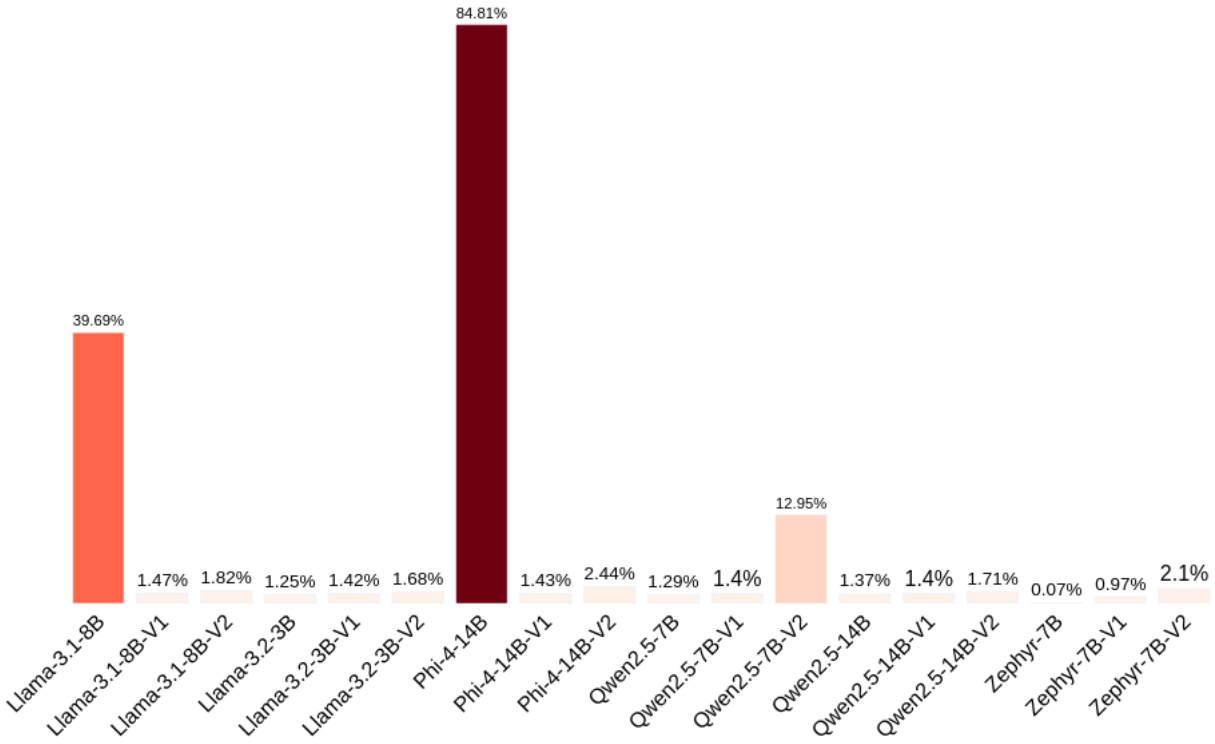
7.2 Resultados Inválidos

No processo de avaliação, considerou-se como resultado inválido toda resposta gerada pelos modelos que não pertencesse ao conjunto de alternativas válidas (A, B, C ou D), conforme o formato exigido pelo conjunto de dados do *MMLU*. Tais respostas indicam que o modelo não interpretou corretamente a estrutura esperada da tarefa, tornando-as inutilizáveis para correção automática.

É importante destacar que, para o cálculo da acurácia, essas respostas inválidas foram computadas como erros, uma vez que o modelo falhou em produzir uma saída aderente ao padrão estipulado. Dessa forma, a resposta era descartada caso não estivesse expressa dentro do

conjunto fechado de alternativas. A Figura 7.1 representa a proporção de resultados inválidos de respostas em cada modelo e seu respectivo versionamento.

Figura 7.1 – Percentual de Respostas Inválidas por Modelo



Fonte: Autora (2025).

De acordo com os dados analisados, observa-se que a maioria dos modelos ajustados com *fine-tuning* (V1 e V2) apresenta taxas de respostas inválidas próximas de zero, o que sugere que o *fine-tuning* contribui significativamente para alinhar a estrutura das respostas ao formato exigido. Por exemplo:

- **Llama-3.1-8B-V1**: apenas 1,47% de respostas inválidas;
- **Qwen2.5-14B-V1**: apenas 1,40%;
- **Zephyr-7B-V1**: menos de 1%.

Por outro lado, modelos na versão base (sem *fine-tuning*) apresentaram índices significativamente maiores de respostas inválidas, o que evidencia a importância do processo de ajuste fino para garantir a conformidade estrutural:

- O **Phi-4-14B** teve taxa de invalidez de 84,81%, o que mostra que, em sua versão original, o modelo não segue o padrão de alternativas definido.

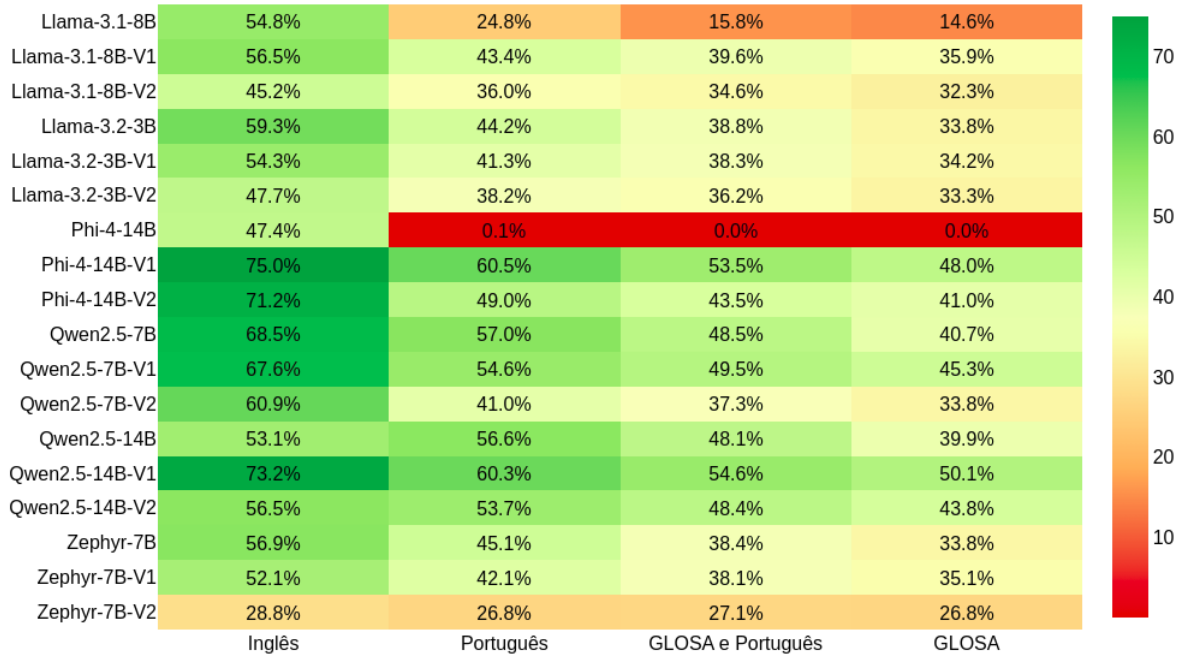
- De forma semelhante, o **Llama-3.1-8B** (base) apresentou 39,68% de respostas inválidas, valor que caiu drasticamente após o *fine-tuning*.
- Um caso específico é o **Qwen2.5-7B-V2**, que mesmo após o ajuste manteve 12,95% de respostas inválidas, o que indica necessidade de refinamento adicional nesta versão.

7.3 Desempenho Geral dos Modelos

O desempenho geral dos modelos frente aos diferentes conjuntos de dados pode ser visualizado no *heatmap* da Figura 7.2, o qual representa as taxas de acurácia (%), isto é, a proporção de respostas corretas obtidas por cada modelo em cada configuração de entrada.

No eixo y, encontram-se os modelos avaliados, tanto na versão original (sem ajustes) quanto nas versões ajustadas via *fine-tuning* (V1 e V2). No eixo x, estão os conjuntos de dados utilizados para teste: inglês original, português traduzido, glosa traduzida e uma mistura de português e glosa. A escala de cores indica a acurácia alcançada em cada combinação modelo–base de dados: tons mais intensos (ou mais claros, conforme a paleta utilizada) correspondem a maior acurácia, enquanto tons menos intensos indicam menor acurácia.

Figura 7.2 – *Heatmap* da Acurácia por Modelo e Tipo de Base de Dados



Fonte: Autora (2025).

De modo geral, os resultados corroboram a tendência antecipada: os modelos apresentam maior acurácia nos conjuntos de dados em Inglês — idioma mais representado nos corpora de pré-treinamento dos LLMs analisados. À medida que se avança para a direita no *heatmap* — passando por contextos em Português padrão, dados mistos (GLOSA + Português) e GLOSA pura — observa-se uma queda progressiva nas taxas de acerto. Tal comportamento indica que o aumento da complexidade linguística e o distanciamento da norma dominante impactam diretamente a performance dos modelos.

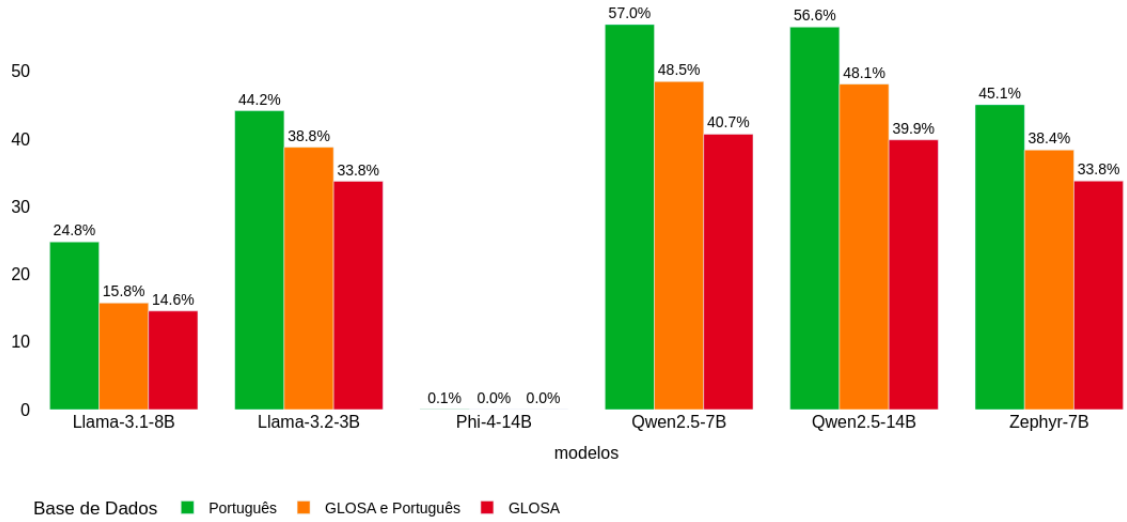
Outro aspecto relevante revelado por essa análise diz respeito ao impacto do *fine-tuning*. De maneira consistente, para praticamente todos os modelos e bases de dados, as versões ajustadas (V1 e V2) apresentaram desempenho superior às versões originais, que não passaram por *fine-tuning*. Essa tendência é facilmente observável ao percorrer cada coluna do *heatmap* de cima para baixo, para cada modelo: na maioria dos casos, os valores de acurácia aumentam conforme se passa da versão base para as versões ajustadas, evidenciando a efetividade do treinamento adicional na adaptação dos modelos às tarefas propostas.

Um exemplo particularmente ilustrativo é o *Phi-4-14B*, cuja versão original apresentou desempenho próximo de zero nos conjuntos em Português e GLOSA. No entanto, após o *fine-tuning*, esse mesmo modelo superou a marca de 40% de acurácia em GLOSA, e, em alguns casos, ultrapassou o desempenho de modelos mais compactos como o *Zephyr-7B*. Esses achados reforçam a ideia central deste trabalho: o *fine-tuning* direcionado é fundamental para que modelos de linguagem generalistas possam lidar de forma eficaz com variações linguísticas restritas ou pouco representadas, como a GLOSA.

7.4 Desempenho nos Contextos Português e GLOSA

A Figura 7.3 apresenta os resultados de desempenho dos modelos em tarefas envolvendo a língua portuguesa, considerando três níveis de variação linguística: Português padrão, GLOSA e uma combinação de GLOSA com Português. Esses testes foram realizados utilizando os modelos em suas versões originais, ou seja, sem aplicação de *fine-tuning*, com o objetivo de evidenciar a robustez de cada arquitetura diante de variações linguísticas que se afastam do Inglês, idioma predominante nos dados de pré-treinamento.

Figura 7.3 – Acurácia por Modelo sem Ajuste Fino para o Contexto Português e GLOSA



Fonte: Autora (2025).

No contexto do Português, observa-se que apenas dois modelos superaram a marca de 50% de acertos: **Qwen2.5-7B** (56,95%) e **Qwen2.5-14B** (56,58%), apresentando equilíbrio entre acertos e erros. O **Zephyr-7B** e o **Llama-3.2-3B** ficaram abaixo desse patamar, com taxas de acerto de 45,09% e 44,19% respectivamente. Já o **Llama-3.1-8B** se destacou negativamente, com apenas 24,81% de acertos e uma taxa elevadíssima de respostas inválidas (46,64%), o que indica sérias dificuldades na geração de respostas adequadas em Português. O **Phi-4-14B** apresentou o pior cenário, com praticamente nenhum acerto (0,05%) e quase todas as saídas inválidas (99,87%).

No cenário combinado de GLOSA e Português, o padrão se repete: apenas os modelos da família Qwen mantiveram desempenho superior a 48% de acertos (**Qwen2.5-7B** com 48,52% e **Qwen2.5-14B** com 48,14%). Já **Zephyr-7B** e **Llama-3.2-3B** ficaram em torno de 38%, enquanto o **Llama-3.1-8B** caiu ainda mais para 15,76%, com impressionantes 57,76% de respostas inválidas. O **Phi-4-14B** mais uma vez foi praticamente inoperante neste contexto, com apenas um acerto e 99,98% de saídas inválidas.

Ao migrar para o cenário totalmente em GLOSA, todos os modelos sofreram quedas ainda mais acentuadas de desempenho. Nenhum ultrapassou 50% de acertos. O **Qwen2.5-7B** foi o mais robusto, atingindo 40,74%, seguido de perto pelo **Qwen2.5-14B** (39,89%). **Zephyr-7B** e **Llama-3.2-3B** mantiveram desempenho modesto, com cerca de 33,8%. O **Llama-3.1-8B** caiu para 14,58% de acertos e novamente exibiu alto índice de respostas inválidas (53,84%). Já o **Phi-4-14B** praticamente não respondeu de forma correta.

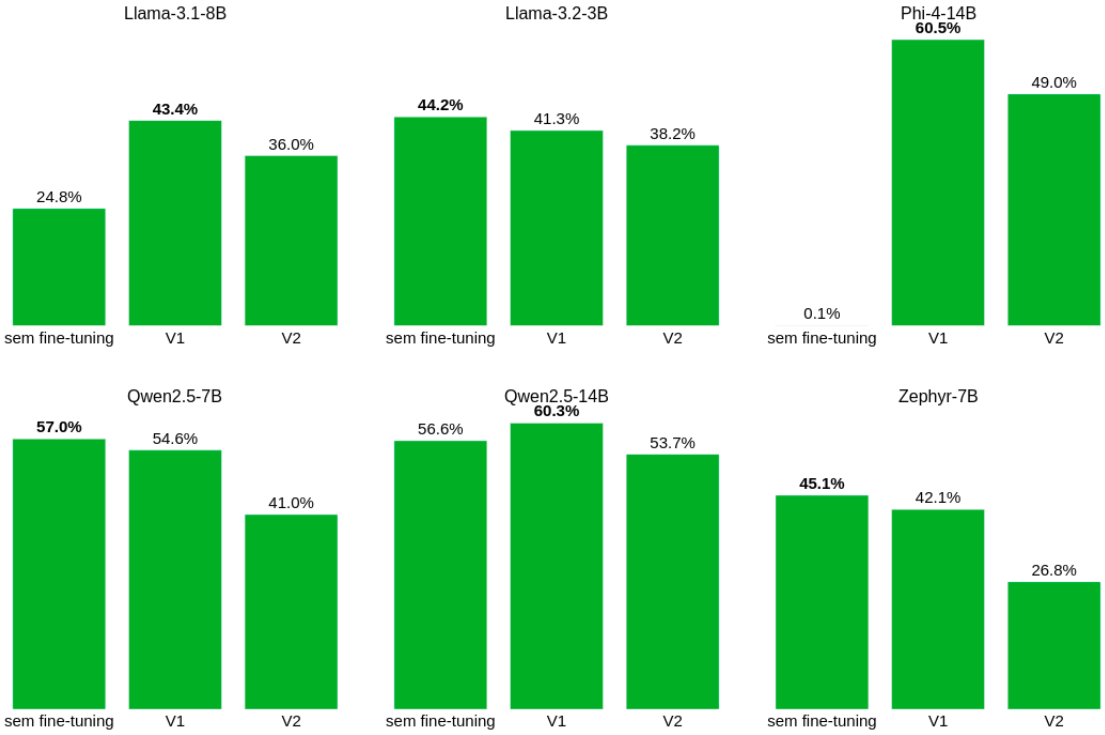
Esses resultados evidenciam que, embora alguns modelos apresentem desempenho razoável em Português padrão e no cenário misto com GLOSA, o uso pleno da variante GLOSA ainda impõe desafios significativos. A queda geral nas taxas de acerto e o aumento das respostas inválidas mostram limitações claras na capacidade dos modelos de lidar com estruturas linguísticas menos convencionais. Isso reforça a importância de estratégias de *fine-tuning* específicas e de dados de treinamento mais adequados para variantes como a GLOSA. Em síntese, adaptar grandes modelos de linguagem a contextos controlados exige intervenções cuidadosas para garantir coerência e precisão.

7.4.1 Impacto do Fine-Tuning

O *fine-tuning* quando bem calibrado, é um meio para adaptar LLM a dialetos, jargões ou variações específicas como a GLOSA. Porém, os resultados obtidos mostram que a qualidade do conjunto de dados de ajuste e a estratégia de *fine-tuning* são determinantes: uma abordagem mal calibrada pode comprometer a acurácia e gerar saídas inválidas, ao invés de melhorar a performance.

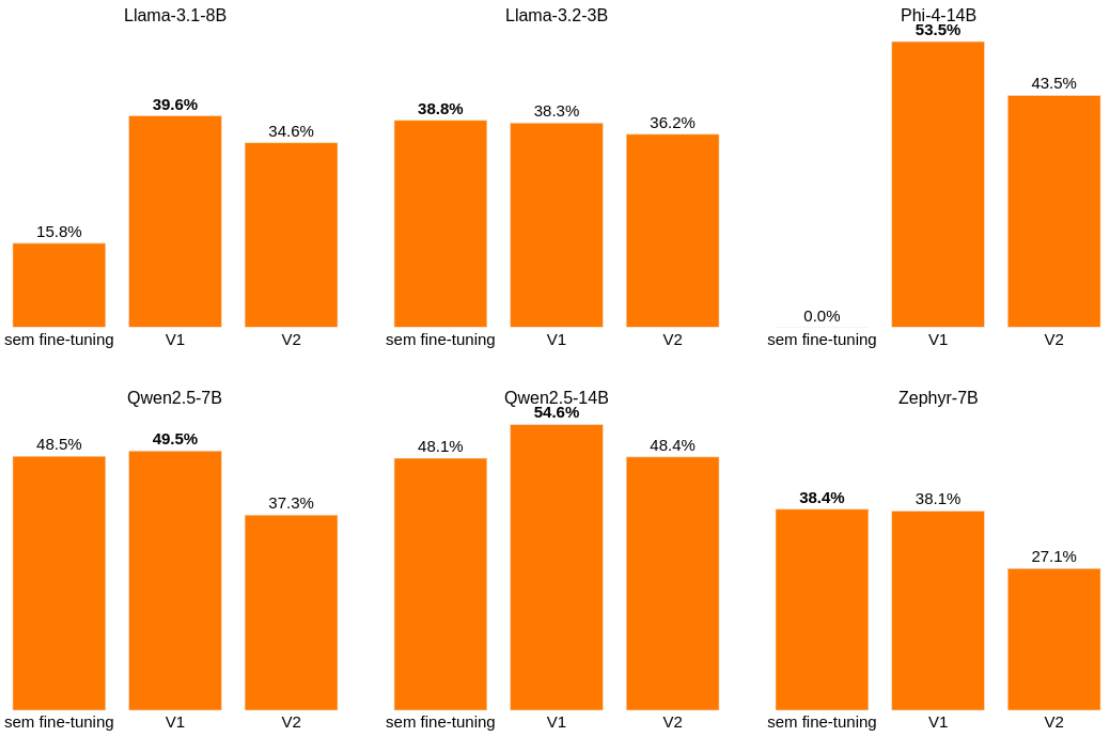
Nesta análise foi comparado o desempenho das versões V0 (modelo original, sem ajuste fino), V1 e V2 de cada modelo, avaliando como o *fine-tuning* específico impactou os percentuais de acertos, erros e saídas inválidas em três contextos: Português padrão, GLOSA e a combinação GLOSA+Português. Essa comparação é essencial para entender o potencial de adaptação dos modelos a variantes linguísticas que fogem do domínio original em Inglês. As Figuras 7.4, 7.5 e 7.6 apresentam esses resultados de forma detalhada.

Figura 7.4 – Comparação das Versões (V0, V1 e V2) em Português



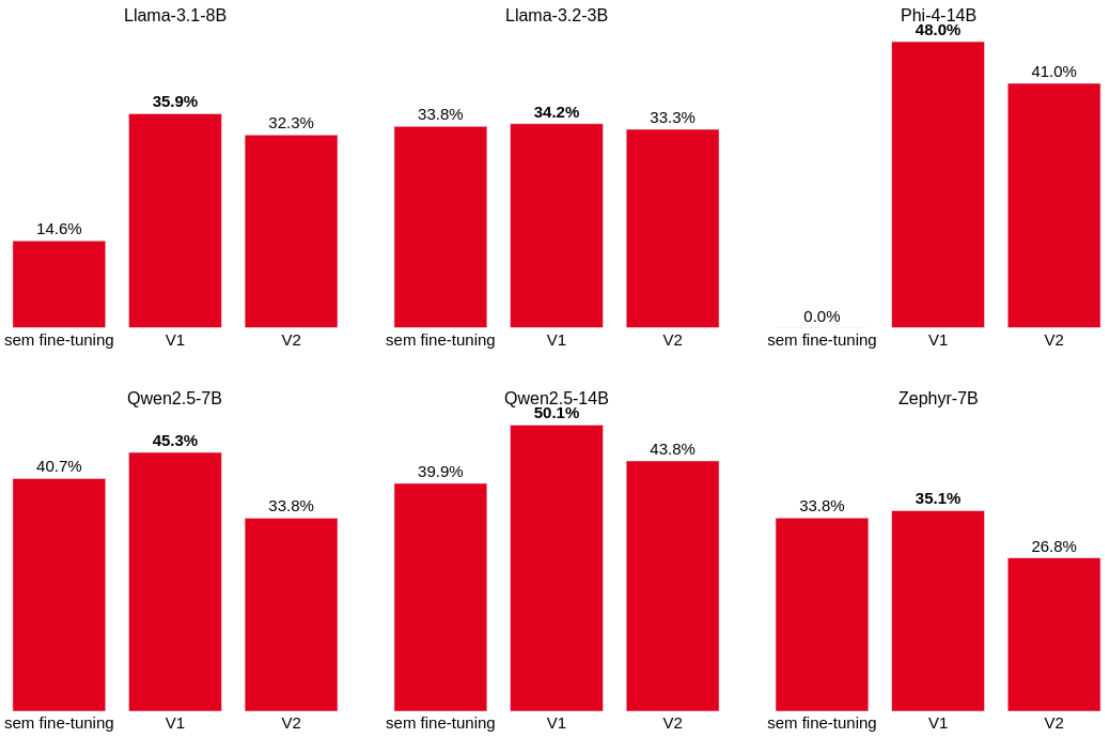
Fonte: Autora (2025).

Figura 7.5 – Comparação das Versões (V0, V1 e V2) em GLOSA e Português



Fonte: Autora (2025).

Figura 7.6 – Comparação das Versões (V0, V1 e V2) em GLOSA



Fonte: Autora (2025).

No Português, observa-se uma tendência geral de queda de desempenho da V1 para a V2 na maioria dos modelos, com exceção de alguns casos pontuais. Por exemplo:

- O **Phi-4-14B** caiu de 60,51% de acertos na V1 para 49,02% na V2, sugerindo perda de precisão após o ajuste.
- O **Qwen2.5-7B** também teve redução expressiva: de 54,59% para 41,03%, acompanhada por um aumento significativo de respostas inválidas (de 3,06% para 16,69%).
- O **Zephyr-7B** reduziu drasticamente de 42,12% (V1) para 26,84% (V2).
- Por outro lado, o **Qwen2.5-14B** manteve desempenho mais estável, com leve redução de 60,30% para 53,73% de acertos, indicando que o modelo sofreu menos com o *fine-tuning* nesse contexto.

Quando analisados em conjunto (**GLOSA + Português**), a queda de desempenho se mantém em quase todos os casos, revelando o desafio adicional de aprender variações combinadas:

- O **Qwen2.5-7B** reduziu de 49,54% (V1) para 37,30% (V2), novamente com alto índice de inválidos na V2 (15,57%).

- O **Phi-4-14B** caiu de 53,54% para 43,46%, mantendo baixo índice de inválidos, mas ainda assim com perda de precisão.
- Já o **Qwen2.5-14B** apresentou o comportamento mais estável, com 54,58% (V1) e 48,39% (V2), sugerindo menor impacto negativo do *fine-tuning* nessa combinação.

No contexto da GLOSA, o impacto do *fine-tuning* se torna mais evidente. Alguns modelos apresentam leve melhora, enquanto outros registram aumento de saídas inválidas:

- O **Phi-4-14B** caiu de 47,97% (V1) para 40,98% (V2), com aumento de inválidos de 0,84% para 2,29%.
- O **Qwen2.5-7B** teve queda de 45,30% (V1) para 33,81% (V2), mas com salto expressivo de inválidos: de 0,83% para 17,67%.
- Modelos como o **Zephyr-7B** e o **Llama-3.1-8B** repetem essa tendência de perda de acurácia, indicando que o *fine-tuning* não capturou adequadamente as nuances da GLOSA isoladamente.

Embora contraintuitiva, essa queda de desempenho pode ser explicada por alguns fatores. O *fine-tuning*, ao ajustar os modelos para tarefas específicas, pode ter causado sobreajuste, reduzindo a capacidade de generalização aprendida na versão V1. Além disso, a GLOSA apresenta nuances linguísticas distintas do Português, que alguns modelos não capturaram adequadamente, resultando em aumento de respostas inválidas. Por fim, limitações no tamanho e na representatividade do conjunto de *fine-tuning* podem ter contribuído para a instabilidade observada. Dessa forma, a redução de desempenho não indica necessariamente falha no *fine-tuning*, mas evidencia os desafios de adaptação a novas representações linguísticas e à manutenção da generalização dos modelos.

7.4.2 Exemplo Prático

Para ilustrar os resultados obtidos, o Quadro 7.2 apresenta exemplos de saídas geradas pelos LLMs durante o processo de inferência, conforme descrito na Figura 6.9 e detalhado na Subseção 6.3.4.

A coluna *Assistent* foi incluída porque, em alguns modelos, após o processo de *fine-tuning*, observou-se que a saída não se limitava à marcação esperada após a *string* ‘resposta”.

Em vez disso, os modelos também acrescentaram a marcação *assistant*, seguida de uma explicação adicional. Esse comportamento sugere que, além de indicar a resposta correta, os modelos passaram a fornecer um breve raciocínio ou justificativa, refletindo um maior nível de detalhamento e profundidade na geração da resposta.

Quadro 7.2 – Exemplos de Respostas dos LLMs — Comparação V0 e V1

Modelo	**V0**	**Assistant V0**	**V1**	**Assistant V1**
Gemma-3	resposta: (a) gord	–	resposta: (a) gord"	–
LLaMA-3.1	resposta: "	–	resposta: (b)	–
LLaMA-3.2	resposta: (a) gordura =	–	resposta: (a) gordura"	–
Phi-4-14B	resposta: "	–	resposta: a	assistant: h
Qwen2.5-7B	resposta: a opção correta	–	resposta: a	assistant: res
Qwen2.5-14B	resposta: (a) gordura"	–	resposta: a	assistant: a
Zephyr-7B	resposta: (d) g"	–	resposta: (d) g"	–

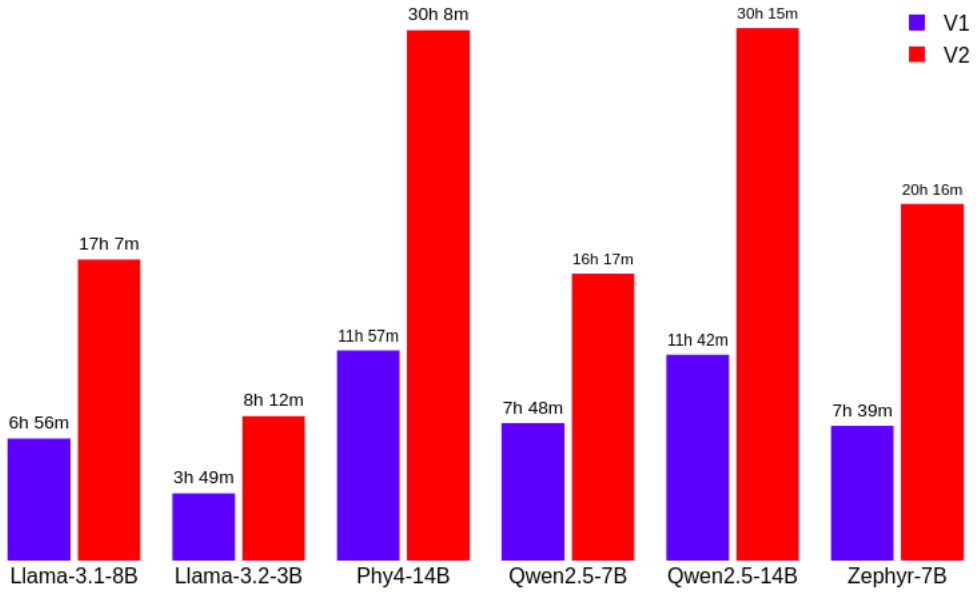
Fonte: Autora (2025).

A análise dos exemplos de respostas geradas pelos modelos nas versões V0 e V1 evidencia que o fine-tuning contribuiu para melhorar a consistência, padronização e interpretabilidade das saídas. Modelos como Gemma-3 e LLaMA-3.2 mantêm o conteúdo correto com pequenas alterações de formatação, enquanto modelos como Qwen2.5-7B e Qwen2.5-14B apresentam respostas mais concisas e estruturadas, indicando que o ajuste favoreceu a organização das saídas em formato de diálogo, especialmente por meio do campo assistant:. Observa-se que modelos maiores, como Phi-4-14B, apresentam maior capacidade de gerar respostas coerentes, embora ainda sejam necessárias intervenções de fine-tuning para alinhar a saída às expectativas linguísticas da GLOSA. Por outro lado, modelos menores ou de porte intermediário, como LLaMA-3.1 e Zephyr, mostraram respostas mais irregulares na V0, evidenciando limitações de pré-treinamento para lidar com construções linguísticas controladas. Além disso, a presença de respostas truncadas ou simplificadas indica a necessidade de curadoria linguística e formatação consistente dos exemplos de treinamento.

7.4.3 Tempo de Treinamento

Além da acurácia, o tempo de treinamento também é um fator relevante na avaliação de desempenho dos modelos, especialmente quando se busca eficiência computacional e viabilidade para experimentação em ambientes com recursos limitados. A Figura 7.7 apresenta o tempo total de treinamento necessário para cada modelo, comparando as versões V1 (em azul) e V2 (em vermelho).

Figura 7.7 – Tempo de Treinamento



Fonte: Autora (2025).

Observa-se que, de forma consistente entre os modelos, a versão V2 demandou significativamente mais tempo de treinamento em comparação com a versão V1. Isso é esperado, pois a V2 foi configurada com foco em robustez e qualidade de resposta, o que implicou em mais épocas, menor taxa de aprendizado e maior uso de contexto. Em contraste, a versão V1 priorizou rapidez e economia de recursos computacionais, resultando em tempos de execução mais curtos.

O Phi-4-14B, um dos modelos mais pesados, levou aproximadamente 11h57min na V1 e 30h08min na V2 — quase o triplo do tempo. O Qwen2.5-14B teve comportamento semelhante: 11h42min na V1 e 30h15min na V2. Já modelos menores, como o Llama-3.2-3B, mostraram tempos mais moderados, com 6h56min na V1 e 17h07min na V2.

Esses dados abordam a relação direta entre a estratégia de treinamento adotada e o custo computacional envolvido. Embora para alguns casos a V2 tende a oferecer melhorias em ter-

mos de qualidade de resposta, sua aplicação deve ser ponderada de acordo com os recursos disponíveis e os objetivos específicos do projeto. Para ciclos rápidos de experimentação e prototipagem, a versão V1 se mostra mais viável. Já para implantações finais ou tarefas mais sensíveis, o custo adicional da V2 pode ser justificado.

7.4.4 Ganhos do Ajuste Fino

Os resultados apresentados no Quadro 7.3 permitem avaliar de forma detalhada se o *fine-tuning* contribui efetivamente para a adaptação dos modelos à variante sintática GLOSA combinada com o Português. Neste Quadro é possível encontrar os modelos bem como a acurácia obtida em cada um, sem a aplicação do *fine-tuning* e com a aplicação do *fine-tuning*. A coluna (%) *Baseline* apresenta os resultados de acurácia quando se é utilizado a base de dados no formato de instrução de perguntas em Glosa e respostas em Português nos modelos originais, isto é, sem a aplicação de *fine-tune*. Já na coluna (%GLOSA+PT) é apresentado a acurácia dos modelos após aplicar o V1 e V2. Por conseguinte, na coluna Ganho (%) é apresentado a diferença percentual de acertos entre o modelo *baseline* e após a realização do *fine-tuning*. O principal objetivo desta análise é verificar se o *fine-tuning* é capaz de elevar o desempenho em tarefas que exigem entendimento de uma estrutura linguística controlada e fora do padrão do idioma majoritário do pré-treinamento.

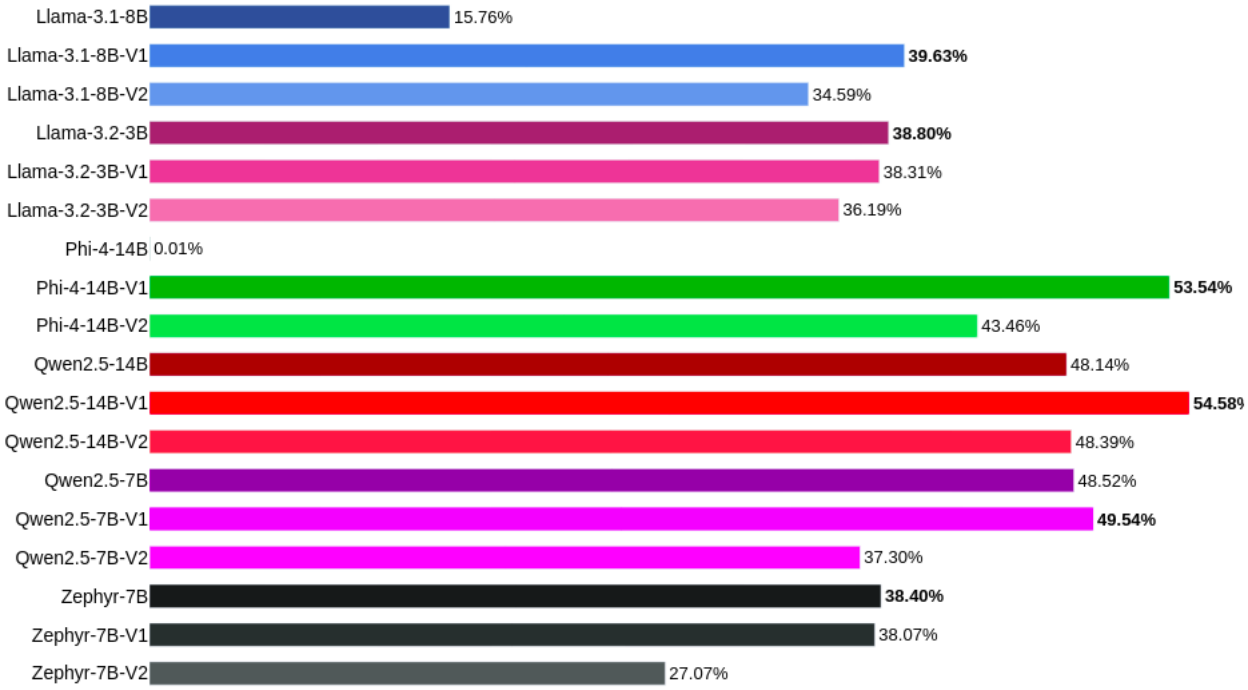
Quadro 7.3 – Comparação de Desempenho no Cenário GLOSA+Português.

Modelo	**(%) Baseline GLOSA + PT**	**(%) GLOSA + PT**	**Ganho (%)**
Llama-3.1-8B-V1	15,76	39,63	+23,87
Llama-3.1-8B-V2	15,76	34,59	+18,83
Llama-3.2-3B-V1	38,80	38,31	–0,49
Llama-3.2-3B-V2	38,80	36,19	–2,61
Qwen2.5-7B-V1	48,52	49,54	+1,02
Qwen2.5-7B-V2	48,52	37,30	–11,22
Qwen2.5-14B-V1	48,14	54,58	+6,44
Qwen2.5-14B-V2	48,14	48,39	+0,25
Phi-4-14B-V1	0,01	53,54	+53,53
Phi-4-14B-V2	0,01	43,46	+43,45
Zephyr-7B-V1	38,40	38,07	–0,33
Zephyr-7B-V2	38,40	27,07	–11,33

Fonte: Autora (2025).

De modo geral, observa-se que os modelos exibem comportamentos distintos após o *fine-tuning*, com alguns apresentando ganhos significativos e outros, perdas de desempenho. Este cenário evidencia que o impacto do *fine-tuning* não é uniforme e depende de fatores como arquitetura, tamanho do modelo e configuração do ajuste. A Figura 7.8 traz o comparativo de acertos dos modelos, dada cada versão.

Figura 7.8 – Acurácia para Perguntas em GLOSA e Respostas em Português por Modelo



Fonte: Autora (2025).

Para o **Phi-4-14B**, o efeito do *fine-tuning* é notável: o modelo praticamente não tinha capacidade de gerar respostas corretas no cenário GLOSA+Português em sua versão original (**Baseline** de apenas 0,01%). Após o *fine-tuning*, o acerto salta para 53,54% (V1) e 43,46% (V2), representando os maiores ganhos absolutos observados entre todos os modelos (+53,53 e +43,45 pontos percentuais, respectivamente). Isso demonstra de forma inequívoca que o *fine-tuning* foi determinante para habilitar o modelo a lidar com a estrutura controlada da GLOSA.

O **Llama-3.1-8B** também mostra avanços importantes. Partindo de uma base bastante limitada (15,76% de acertos), o modelo alcança 39,63% (V1) e 34,59% (V2) após o ajuste, evidenciando ganhos de +23,87 p.p. e +18,83 p.p. Esses resultados reforçam que, mesmo modelos que inicialmente apresentam dificuldades em entender a GLOSA podem se beneficiar substancialmente de um ajuste dedicado.

Outro destaque é o **Qwen2.5-14B**, que demonstra maior robustez: seu ganho na V1 é de +6,44 p.p., elevando o desempenho de 48,14% para 54,58%. Na V2, o modelo mantém praticamente o mesmo patamar, com 48,39%. Essa estabilidade sugere que modelos maiores e com maior capacidade de parâmetros podem absorver melhor o *fine-tuning* sem sofrer perdas substanciais.

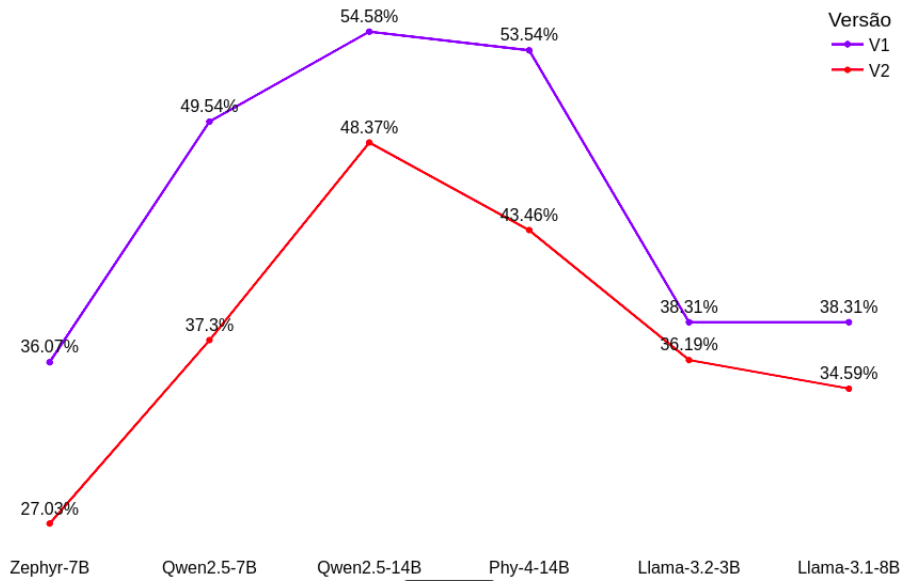
Por outro lado, nem todos os modelos respondem positivamente ao *fine-tuning*. O **Qwen2.5-7B**, por exemplo, parte de um desempenho já relativamente alto no cenário Baseline (48,52%), mas apresenta um ganho modesto na versão V1 (+1,02 p.p.) e uma queda acentuada na versão V2 (–11,22 p.p.). Isso sugere que, para este modelo específico, o *fine-tuning* pode ter induzido uma sobre-especialização ou até degradado aspectos gerais do conhecimento adquirido, prejudicando a coerência das respostas.

De forma semelhante, o **Zephyr-7B** mostra estabilidade praticamente nula na V1 (–0,33 p.p.) e uma queda significativa na V2 (–11,33 p.p.). Esses resultados indicam que o *fine-tuning* precisa ser cuidadosamente calibrado para modelos menores ou mais sensíveis a variações no conjunto de dados, a fim de evitar perda de generalização.

Já o **Llama-3.2-3B** apresenta leve queda de desempenho em ambas as versões ajustadas (–0,49 p.p. na V1 e –2,61 p.p. na V2), o que sugere que o *fine-tuning* aplicado não foi suficiente para ampliar sua competência no contexto da GLOSA.

Na Figura 7.9, observa-se que para todos os modelos, a V1 apresentou taxas de acerto superiores à V2, mesmo com tempo de treinamento significativamente menor, conforme apresentado anteriormente na Figura 7.7. Isso indica que, embora a versão V2 tenha demandado mais tempo e recursos computacionais, a versão V1 se mostrou mais eficiente em termos de relação custo-benefício. Ou seja, o ganho obtido na V2 não justifica o esforço adicional de treinamento.

Figura 7.9 – Acurácia dos Modelos para Base de Dados Perguntas em GLOSA e Respostas em Português



Fonte: Autora (2025).

Os resultados apresentados demonstram que, de maneira geral, o *fine-tuning* é uma etapa que pode trazer benefícios relevantes para os modelos de linguagem, especialmente em contextos que envolvem variantes linguísticas específicas, como a GLOSA. No entanto, o impacto do ajuste fino varia significativamente entre os modelos e configurações. A versão V1 do *fine-tuning* se destacou por alcançar melhorias de desempenho em diversas arquiteturas, com custo computacional substancialmente menor que a versão V2, o que a torna uma escolha mais viável em contextos de experimentação rápida e limitação de recursos.

Considerando o equilíbrio entre acurácia, aderência ao formato de resposta (baixa taxa de erro), tempo de treinamento e robustez diante da variação linguística, o modelo **Llama-3.2-3B-V1** emerge como o mais indicado para aplicações com restrição de recursos, apresentando bom desempenho em tarefas em GLOSA e Português, além de baixo tempo de ajuste. Para cenários em que há maior disponibilidade computacional e foco em desempenho em tarefas mais complexas, o **Qwen2.5-14B-V1** é o mais promissor, superando outros modelos em acurácia com taxas mínimas de respostas inválidas.

7.5 Análise de Erros

Durante a avaliação dos modelos de linguagem no cenário GLOSA+Português, observou-se uma taxa de erro significativa, mesmo após a aplicação do ajuste fino. Isso reforça a complexidade da tarefa e a sensibilidade dos modelos ao tipo e estrutura dos dados de entrada.

De modo geral, alguns padrões se destacam na origem desses erros. Um dos mais consistentes é a associação entre erros e comprimentos maiores dos *prompts*. Os modelos tendem a apresentar maior índice de falhas quando os enunciados são mais longos, o que pode indicar dificuldades na retenção de contexto ou na coerência das respostas geradas em sequências extensas.

O Quadro 7.4 apresenta um resumo estatístico do comprimento médio dos *prompts* de entrada — medido em número de *tokens*, que corresponde à quantidade de unidades textuais processadas pelo modelo — e das saídas geradas. Os valores foram calculados apenas para os casos em que os modelos apresentaram erros, permitindo comparar seu desempenho na configuração original (*baseline*) e após o ajuste por *fine-tuning* (V1). Observa-se que os modelos, tanto antes quanto depois do *fine-tuning*, apresentam médias de comprimento de *prompt* acima de 500 *tokens* nos casos de erro.

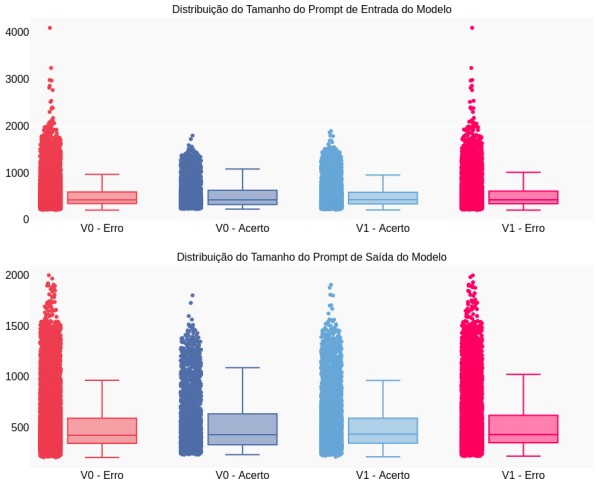
Quadro 7.4 – Comparação dos Erros Antes e Depois do *fine-tuning* no Cenário Perguntas em GLOSA e Respostas em Português.

Modelos	***fine-tuning***	**Tamanho médio de entrada**	**Tamanho médio de saída**	**Total de instâncias**
Llama-3.1-8B	Não	516,13	515,64	12.678
Llama-3.1-8B	Sim	527,53	533,62	9.085
Llama-3.2-3B	Não	524,89	531,71	9.211
Llama-3.2-3B	Sim	523,32	528,65	9.285
Phi-4-14B	Não	517,88	515,08	15.049
Phi-4-14B	Sim	530,67	539,87	6.992
Qwen2.5-7B	Não	529,72	535,44	7.747
Qwen2.5-7B	Sim	534,58	545,23	7.594
Qwen2.5-14B	Não	529,75	540,88	7.805
Qwen2.5-14B	Sim	537,68	547,19	6.835
Zephyr-7B	Não	518,95	518,19	9.271
Zephyr-7B	Sim	520,36	519,22	9.321

Fonte: Autora (2025).

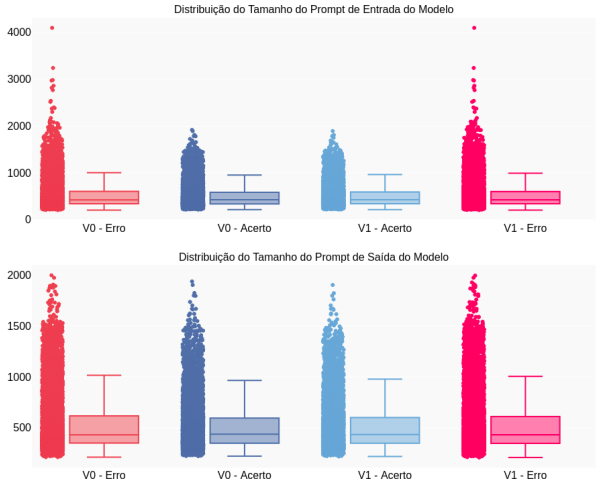
Em muitos casos, os *prompts* com erro são mais longos do que os que resultaram em acertos. Para investigar esse padrão mais profundamente, foram gerados *box plots* para cada modelo, representando a distribuição do tamanho dos *prompts* nos casos de acerto e erro. Esses gráficos, apresentados nas Figuras 7.10 7.11 7.12 7.13 7.14 7.15, reforçam a ideia de que contextos mais longos estão mais frequentemente associados a respostas incorretas.

Figura 7.10 – Distribuição dos Erros e Acertos por Tamanho de Prompt - LLaMA 3.1



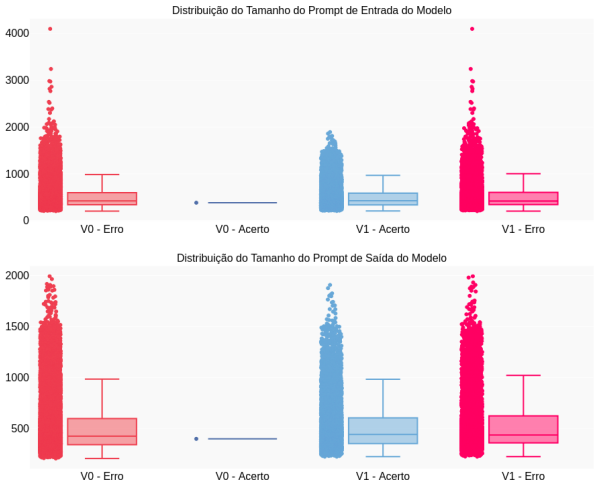
Fonte: Autora (2025).

Figura 7.11 – Distribuição dos Erros e Acertos por Tamanho de Prompt - LLaMA 3.2



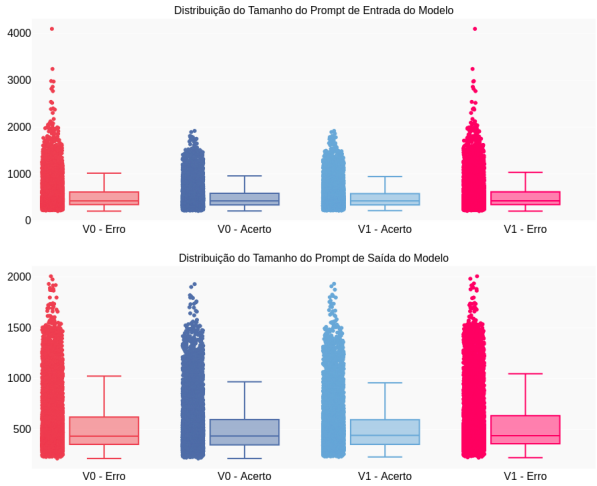
Fonte: Autora (2025).

Figura 7.12 – Distribuição dos Erros e Acertos por Tamanho de Prompt - Phi-4-14B



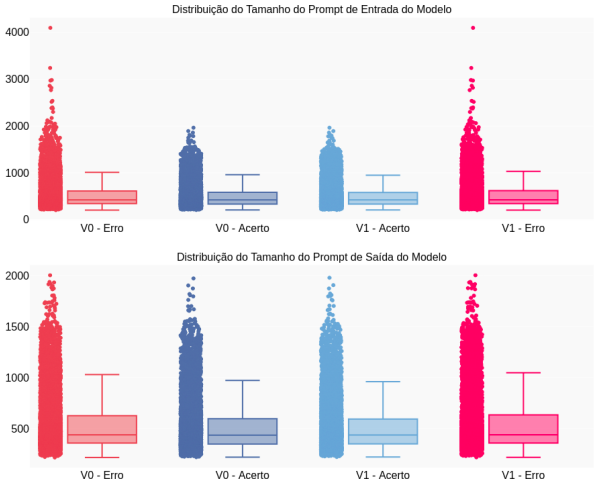
Fonte: Autora (2025).

Figura 7.13 – Distribuição dos Erros e Acertos por Tamanho de Prompt - Qwen2.5-7B



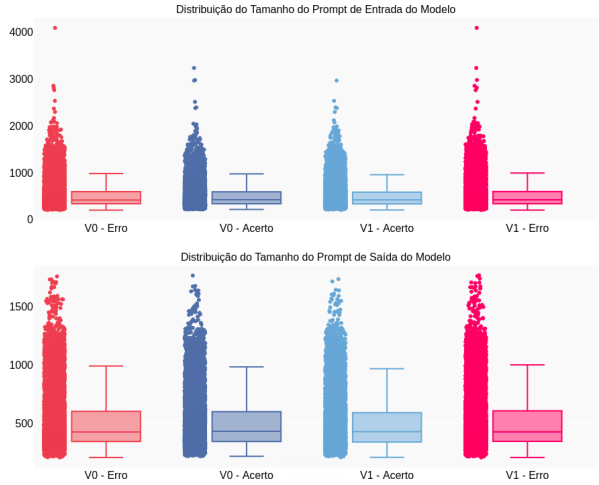
Fonte: Autora (2025).

Figura 7.14 – Distribuição dos Erros e Acertos por Tamanho de Prompt - Qwen2.5-14B



Fonte: Autora (2025).

Figura 7.15 – Distribuição dos Erros e Acertos por Tamanho de Prompt - Zephyr-7B



Fonte: Autora (2025).

Embora a associação entre erros e comprimento seja clara, ainda não se sabe com precisão o que motiva esse comportamento. Pode haver múltiplas causas, como limitação de contexto interno dos modelos, qualidade da tradução GLOSA em exemplos longos, ou falhas na generalização em sequências complexas.

A seguir, apresentam-se os Quadros que detalham a distribuição dos erros cometidos pelos diferentes modelos em função dos assuntos das perguntas. O objetivo desta análise é identificar se há tópicos específicos nos quais os modelos apresentam maior dificuldade de interpretação.

O Quadro 7.5 mostra os top 10 erros por assunto para os modelos da família LLaMA, onde se observa que os maiores números de erros ocorrem em questões relacionadas a Direito Profissional e Cenários Morais. Os modelos com *fine-tuning* na V1 tendem a apresentar menor quantidade de erros, indicando melhorias consistentes.

Quadro 7.5 – Percentual de Erros por Assunto — Modelos LLaMA.

Assunto	**LLaMA 3.1 8B-V0**	**LLaMA 3.1 8B-V1**	**LLaMA 3.2 3B-V0**	**LLaMA 3.2 3B-V1**
Direito Profissional	9,18%	7,73%	7,71%	7,53%
Cenários Morais	6,49%	4,81%	4,76%	4,72%
Assuntos Diversos	3,83%	3,00%	3,17%	3,20%
Psicologia Profissional	3,79%	2,70%	2,80%	2,77%
Psicologia Ensino Médio	3,07%	2,00%	2,04%	2,05%
Macroeconomia Ensino Médio	2,59%	1,77%	1,85%	1,96%
Matemática Fundamental	2,42%	1,73%	1,65%	1,82%
Disputas Morais	2,11%	1,55%	1,48%	1,46%
Filosofia	1,98%	1,35%	1,33%	1,34%
Contabilidade Profissional	1,93%	1,40%	1,43%	1,41%

Fonte: Autora (2025).

No Quadro 7.6, são os top 10 erros referente aos modelos Qwen, o padrão é similar: Direito Profissional novamente destaca-se como o assunto com maior dificuldade, seguido por Cenários Morais e Assuntos Diversos. Nota-se também a redução de erros nas V1, embora algumas categorias menos frequentes, como Filosofia, tenham valores zerados, indicando que eles obtiveram 100% de acertos nessa categoria .

Quadro 7.6 – Percentual de Erros por Assunto — Modelos Qwen.

Assunto	**Qwen 2.5 14B-V0**	**Qwen 2.5 14B-V1**	**Qwen 2.5 7B-V0**	**Qwen 2.5 7B-V1**
Direito Profissional	7,33%	6,99%	7,23%	7,23%
Cenários Morais	4,77%	4,68%	4,80%	4,91%
Assuntos Diversos	2,24%	2,04%	2,38%	2,28%
Psicologia Profissional	2,23%	2,02%	2,37%	2,28%
Psicologia Ensino Médio	1,47%	1,11%	1,53%	1,38%
Macroeconomia Ensino Médio	1,32%	1,04%	1,36%	1,28%
Matemática Fundamental	1,39%	1,10%	1,37%	1,28%
Disputas Morais	1,37%	1,21%	1,28%	1,29%
Filosofia	1,07%	0,95%	0,00%	0,00%
Contabilidade Profissional	1,28%	1,21%	1,33%	1,29%

Fonte: Autora (2025).

Por fim, o Quadro 7.7 apresenta os top 10 erros por assuntos, referentes aos modelos Zephyr e Phi. Novamente, a área de Direito Profissional apresenta a maior quantidade de erros, especialmente no modelo Phi-4 14B-V0. Isso provavelmente se deve ao fato de essas questões envolverem um vocabulário menos conhecido e disseminado, cuja complexidade aumenta a dificuldade de interpretação nesse tema. Além disso, percebe-se que os modelos Phi tendem a apresentar maiores valores absolutos de erros, sugerindo maior dificuldade nesses temas ou diferenças no conjunto de dados.

Quadro 7.7 – Percentual de Erros por Assunto — Modelos Zephyr e Phi.

Assunto	**Zephyr 7B-V0**	**Zephyr 7B-V1**	**Phi 4 14B-V0**	**Phi 4 14B-V1**
Direito Profissional	7,45%	7,52%	11,01%	6,78%
Cenários Morais	4,97%	4,87%	6,57%	4,90%
Assuntos Diversos	3,02%	2,98%	5,72%	2,04%
Psicologia Profissional	2,96%	2,93%	4,41%	2,09%
Psicologia Ensino Médio	2,02%	2,01%	3,97%	1,29%
Macroeconomia Ensino Médio	1,75%	1,68%	2,85%	1,17%
Matemática Fundamental	1,70%	1,69%	2,51%	1,22%
Disputas Morais	1,57%	1,60%	2,53%	1,12%
Filosofia	1,49%	1,35%	0,00%	0,00%
Contabilidade Profissional	1,43%	1,47%	2,05%	1,26%

Fonte: Autora (2025).

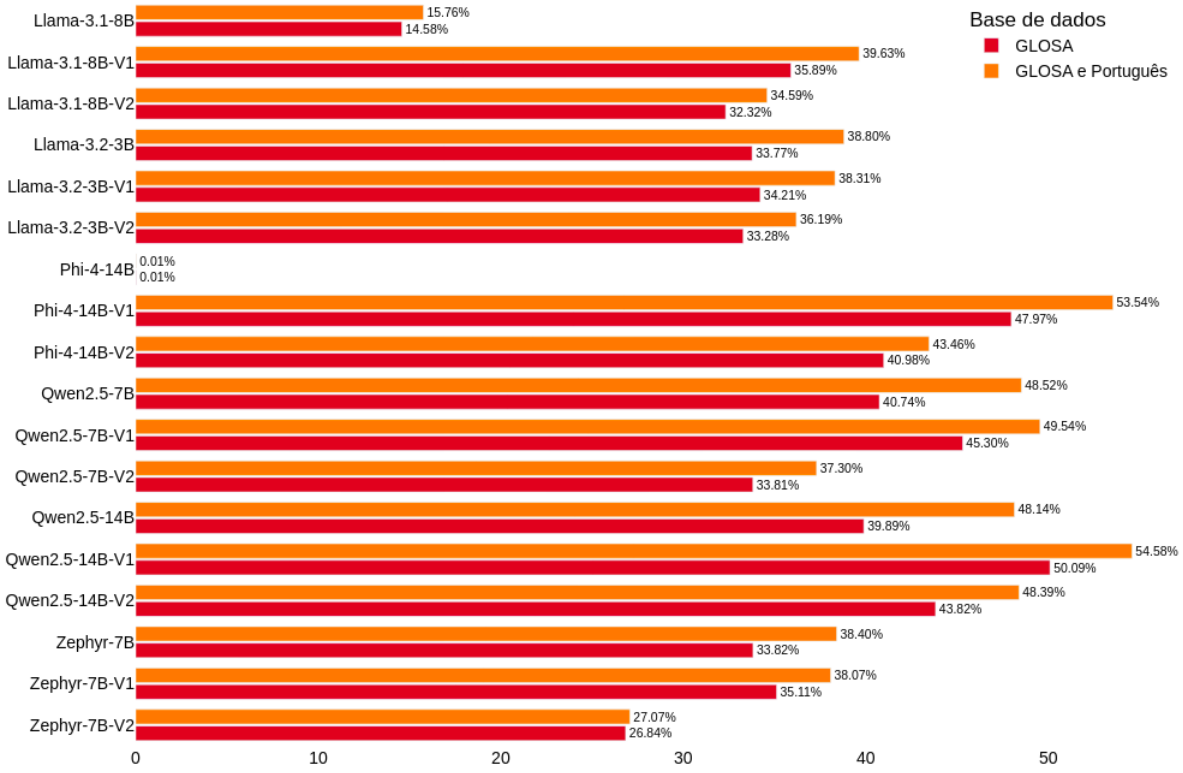
Esta análise indica que, independentemente do modelo, os assuntos relacionados ao Direito e Cenários Morais são os que mais desafiam os modelos, seguido por temas mais diversos e de menor frequência. O *fine-tuning* tende a reduzir os erros de forma consistente, mas não elimina completamente as dificuldades em certos assuntos.

7.6 Limitações e Desafios para a Variante GLOSA

A Figura 7.16 ilustra os principais desafios enfrentados pelos modelos na interpretação da variante linguística GLOSA. Os resultados indicam que o *fine-tuning* com dados formulados em GLOSA melhora parcialmente a capacidade dos LLMs de responder corretamente quando os enunciados estão nessa variante e as alternativas de resposta permanecem em português. Contudo, o desempenho ainda é limitado nos casos em que tanto a pergunta quanto as respostas

estão integralmente formuladas em GLOSA. Esse cenário evidencia a complexidade estrutural da GLOSA para modelos de linguagem treinados predominantemente em dados padrão e aponta para a necessidade de avanços metodológicos. Entre as possíveis estratégias futuras, poderia-se considerar o treinamento de modelos do zero utilizando grandes volumes de dados em GLOSA ou a geração automatizada de corpora em larga escala por meio de *parsers* ou outras ferramentas de conversão linguística. Dessa forma, este trabalho representa um ponto de partida relevante para o desenvolvimento de abordagens mais eficazes no tratamento computacional de variantes linguísticas pouco representadas.

Figura 7.16 – Acurácia dos Modelos Usando GLOSA vs Perguntas em Glosa e Respostas em Português



Fonte: Autora (2025).

Para ilustrar ainda mais essa lacuna, o Quadro 7.8 apresenta a diferença percentual de desempenho dos modelos ao interpretarem entradas em GLOSA combinadas com Português, em comparação com a interpretação das questões exclusivamente em GLOSA. Observa-se que, embora o *fine-tuning* tenha elevado significativamente a capacidade de alguns modelos, como no caso do **Phi-4-14B**, que parte de praticamente zero entendimento para resultados expressivos, o ganho não é uniforme entre todas as arquiteturas. Em muitos casos, mesmo após o *fine-tuning*, ainda há perdas de desempenho ao lidar com a variante linguística pura.

Quadro 7.8 – Comparação de desempenho no cenário Perguntas em GLOSA e Respostas em Português.

Modelos	% GLOSA+PT	% GLOSA	Diferença (pp)
Llama-3.1-8B	15,76	14,58	+1,18
Llama-3.1-8B-V1	39,63	35,89	+3,74
Llama-3.1-8B-V2	34,59	32,32	+2,27
Llama-3.2-3B	38,80	33,77	+5,03
Llama-3.2-3B-V1	38,31	34,21	+4,10
Llama-3.2-3B-V2	36,19	33,28	+2,91
Phi-4-14B	0,01	0,01	0,00
Phi-4-14B-V1	53,54	47,97	+5,57
Phi-4-14B-V2	43,46	40,98	+2,48
Qwen2.5-7B	48,52	40,74	+7,78
Qwen2.5-7B-V1	49,54	45,30	+4,24
Qwen2.5-7B-V2	37,30	33,81	+3,49
Qwen2.5-14B	48,14	39,89	+8,25
Qwen2.5-14B-V1	54,58	50,09	+4,49
Qwen2.5-14B-V2	48,39	43,82	+4,57
Zephyr-7B	38,40	33,82	+4,58
Zephyr-7B-V1	38,07	35,11	+2,96
Zephyr-7B-V2	27,07	26,84	+0,23

Fonte: Elaboração própria.

No conjunto, a análise confirma que o *fine-tuning* pode ser benéfico para habilitar a interpretação da GLOSA, principalmente em modelos que partem de um desempenho muito baixo ou inexistente na variante controlada. Porém, os resultados também alertam que o ajuste não garante melhora automática. É fundamental balancear a estratégia de treinamento (número de épocas, taxa de aprendizado, tamanho do contexto) para cada arquitetura, além de explorar diferentes combinações de dados bilíngues e monolíngues.

7.7 Análises Adicionais

Além das análises iniciais de desempenho sobre a base original do MMLU em Inglês, esta seção apresenta investigações complementares com foco em aspectos específicos do comportamento dos modelos. Inicialmente, avalia-se o desempenho dos LLMs em suas versões

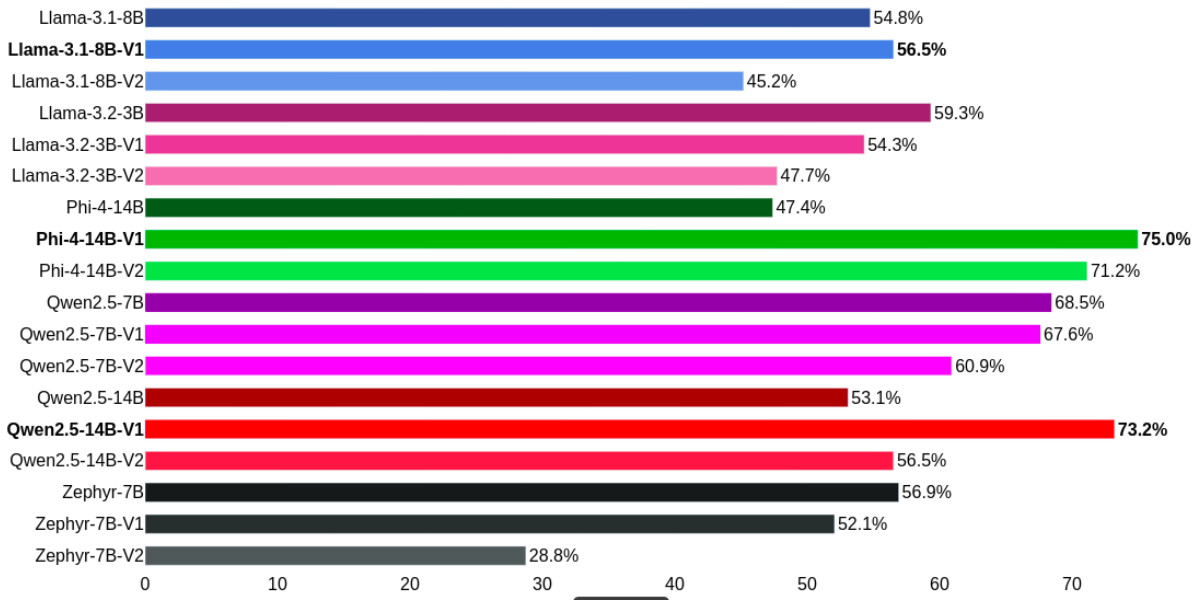
baseline, sem qualquer *fine-tuning*, comparando com as versões ajustadas nas configurações V1 e V2. Em seguida, explora-se a capacidade multilíngue das redes, analisando o impacto da tradução dos dados para o Português sobre a acurácia e consistência das respostas.

Essas análises adicionais permitem identificar não apenas a eficácia dos ajustes de *fine-tuning*, mas também a robustez intrínseca dos modelos frente a variações de idioma e formato de dados, oferecendo um panorama mais detalhado sobre suas limitações e potencialidades em cenários práticos de uso.

7.7.1 Desempenho dos Modelos - Base de dados *Baseline*

Os resultados apresentados na Figura 7.17 estabelece a análise de desempenho dos modelos de acordo com a base de dados do MMLU em seu estado original, isto é, em Inglês, sem a tradução em Português e pré-processamento dos dados. Essa análise contempla tando os modelos sem passar pelo *fine-tuning* quanto para as versões 1 e 2 após *fine-tuning*.

Figura 7.17 – Acurácia dos Modelos Usando a Base de Dados do MMLU em Inglês



Fonte: Autora (2025).

Em suas versões originais, os modelos apresentaram desempenho variado, com acurácia entre aproximadamente 47% (Phi-4-14B) e 68% (Qwen2.5-7B). Esses resultados refletem a capacidade inerente de cada modelo em lidar com tarefas de múltipla escolha no idioma Inglês, para o qual todos foram extensivamente treinados.

Apesar de o *fine-tuning* realizado neste trabalho não ter como objetivo específico a melhoria de desempenho no MMLU, mas sim a adaptação dos modelos ao formato conversacional — conforme a estrutura dos dados Alpaca —, observou-se, em alguns casos, melhora nas taxas de acerto após o ajuste. Por exemplo, o Phi-4-14B, após o *fine-tuning* na configuração V1, elevou sua acurácia de 47% para aproximadamente 75%, e reduziu substancialmente o percentual de respostas inválidas, indicando maior aderência ao formato esperado de resposta.

É importante ressaltar, no entanto, que tais ganhos não são esperados de forma generalizada, dado que os dados utilizados no *fine-tuning* (Alpaca) não foram construídos para treinar modelos em tarefas de múltipla escolha como o MMLU. Esses dados são voltados à construção de respostas abertas no estilo de assistentes conversacionais, o que pode não refletir diretamente em melhorias de desempenho em *benchmarks* formais.

Adicionalmente, a configuração V2 de *fine-tuning*, com foco em respostas mais completas e refinadas, produziu resultados menos homogêneos. Enquanto alguns modelos, como o Qwen2.5-7B e o próprio Phi-4-14B, mantiveram bons níveis de acurácia, outros — como o Zephyr-7B e o Llama-3.2-3B — apresentaram quedas perceptíveis. Esses resultados sugerem que, embora o *fine-tuning* possa melhorar a robustez do modelo em determinados contextos, ele pode também introduzir ruído ou desajustes quando o domínio da tarefa avaliada diverge do domínio dos dados utilizados no treinamento adicional.

7.7.2 Comparação de Desempenho Português vs Inglês

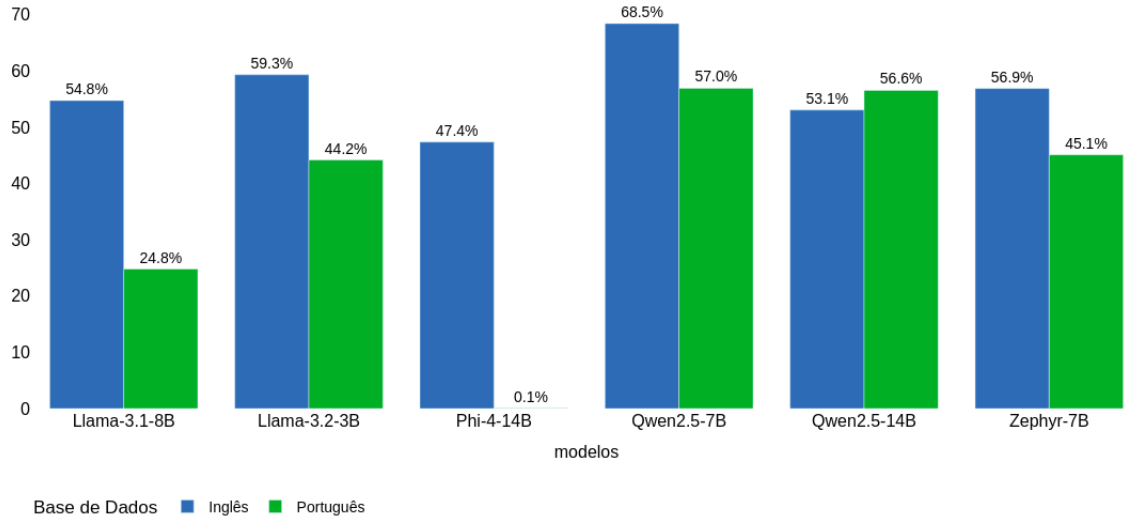
Para avaliar a capacidade multilíngue dos modelos, analisou-se a diferença de desempenho entre o *dataset* em Inglês — língua dominante no pré-treinamento — e o Português padrão.

Conforme ilustrado na Figura 7.18 e detalhado no Quadro 7.9, a maioria dos modelos apresentou queda significativa na acurácia ao migrar para o Português. O modelo **Phi-4-14B** destacou-se pela maior perda absoluta, passando de 47,39% no Inglês para apenas 0,05% no Português, redução de aproximadamente 47 pontos percentuais, além de registrar a maior taxa de respostas inválidas, próxima a 100%.

O **Llama-3.1-8B** também evidenciou queda expressiva, de quase 30 pontos percentuais, corroborando a ideia de que, sem *fine-tuning*, sua capacidade de generalização para o Português é limitada. Modelos como **Zephyr-7B**, **Llama-3.2-3B** e **Qwen2.5-7B** apresentaram quedas moderadas, entre 11 e 15 pontos percentuais, indicando dependência do idioma predominante

no pré-treinamento. Em contrapartida, o **Qwen2.5-14B** apresentou leve ganho de desempenho no Português, sugerindo maior robustez multilíngue intrínseca.

Figura 7.18 – Acurácia dos Modelos sem *fine-tuning* para as Bases de Dados em Inglês e em Português



Fonte: Autora (2025).

Quadro 7.9 – Diferença Percentual de Acurácia entre Datasets em Inglês e Português.

Modelo	**Inglês (%)**	**Português (%)**	**Diferença (p.p.)**
Zephyr-7B	56,91	45,09	–11,82
LLaMA-3.2-3B	59,34	44,19	–15,15
LLaMA-3.1-8B	54,76	24,81	–29,95
Qwen2.5-7B	68,45	56,95	–11,50
Qwen2.5-14B	53,08	56,58	+3,50
Phi-4-14B	47,39	0,05	–47,34

Fonte: Autora (2025).

8 CONCLUSÃO

Este trabalho teve como objetivo principal investigar a capacidade de adaptação de LLMs à variante linguística GLOSA, de modo a subsidiar o desenvolvimento de *chatbots* mais acessíveis e eficazes para a comunidade surda, que utiliza o Português como segunda língua. Para isso, foram conduzidos experimentos comparativos utilizando múltiplas arquiteturas de modelos de linguagem abertos, variando-se o tamanho, as versões e as estratégias de *fine-tuning* aplicadas.

Os resultados confirmam o padrão já observado em estudos multilíngues: os LLMs, mesmo pré-treinados em corpus diversos, têm seu melhor desempenho no inglês — servindo como linha de base de desempenho. Ao migrar para o Português e, principalmente, para a GLOSA, há uma perda sistemática de desempenho. Essa lacuna é ainda mais evidente em construções sintáticas envolvendo apenas a GLOSA, ou seja, com perguntas e respostas em GLOSA.

Com base nos experimentos realizados, foi possível investigar o potencial de adaptação de LLMs à variante linguística GLOSA, especialmente em interações que também envolvem o Português padrão. O estudo avaliou diversas arquiteturas abertas, em múltiplos tamanhos, com duas estratégias distintas de *fine-tuning*. Os modelos foram testados em diferentes contextos linguísticos, variando entre dados em Inglês, Português, GLOSA e mistos.

Os resultados indicam que o *fine-tuning* direcionado pode, sim, ampliar a capacidade de generalização dos modelos, embora seu impacto varie significativamente conforme a arquitetura e o tamanho do modelo. Em geral, modelos maiores, com pré-treinamento multilíngue robusto, apresentaram maior capacidade de adaptação, especialmente quando expostos a dados mistos de GLOSA e Português. Por outro lado, observou-se que ajustes muito intensivos ou desalinhados podem levar a perdas de desempenho, sugerindo efeitos como sobreajuste ou má configuração de hiperparâmetros. A comparação entre versões de ajuste (V1 e V2) revelou que os ganhos não são uniformes, e que diferentes estratégias podem produzir efeitos bastante distintos. Isso reforça a dificuldade de se determinar, de forma antecipada, qual configuração é ideal, especialmente em experimentos de grande porte, nos quais a busca exaustiva por combinações ótimas se torna inviável.

A partir do conjunto de evidências obtido, conclui-se que o *fine-tuning* com dados cuidadosamente construídos, especialmente em formatos que combinem GLOSA e Português padrão,

representa uma abordagem eficaz para melhorar a interpretabilidade dos modelos em contextos linguísticos não convencionais. Contudo, tais ajustes não garantem ganhos automáticos, sendo fundamental o controle do processo de treinamento.

Dentre os modelos avaliados, destacam-se aqueles com maior capacidade de representação e cobertura linguística, como o Phi-4-14B e o Qwen2.5-14B, que demonstraram desempenho sólido após o ajuste. O Llama-3.1-8B também mostrou boa resposta ao treinamento, mesmo sendo de porte intermediário. Com base nos resultados, esses modelos são os mais indicados para uso em sistemas de conversação em forma de *chatbots* voltados a contextos multilíngues ou a línguas com sintaxe controlada, como a GLOSA.

Por fim, este trabalho contribui com um conjunto inédito de dados em GLOSA e Português, um pipeline de experimentação reprodutível e uma análise empírica que reforça a importância da curadoria linguística, da validação contínua e da adequação arquitetural no uso de LLMs em cenários linguísticos específicos. Os códigos dos experimentos podem ser encontrados no Github [1] e o artigo publicado como parte desse trabalho pode ser encontrado na Biblioteca da Sociedade Brasileira de Computação [2].

Apesar dos avanços alcançados neste trabalho, o principal desafio que permanece é o de desenvolver modelos capazes de compreender construções linguísticas que simulam a forma como a comunidade surda se comunica por escrito, especialmente quando essa comunicação se distancia do padrão do Português formal. A GLOSA, neste contexto, foi utilizada como um recurso de representação controlada, buscando aproximar a estrutura textual daquilo que é produzido por surdos em interações escritas, e não como uma variedade linguística amplamente adotada por uma comunidade específica.

Dessa forma, os próximos passos devem concentrar-se na ampliação e diversificação dos dados de treinamento, incorporando contextos mais variados e realistas. Além disso, será necessário investigar estratégias de adaptação mais refinadas, como o uso de aprendizado incremental e ajustes contínuos, sempre com foco em evitar sobreajustes e manter a generalização dos modelos. Uma abordagem alternativa promissora seria encapsular o LLM por meio de um pré-processamento que realize o mapeamento de GLOSA para o português padrão antes da entrada no modelo, possivelmente utilizando um *parser*. Essa estratégia poderia reduzir o esforço de ajuste direto do LLM e facilitar a interpretação de variantes linguísticas pouco representadas.

[1] Disponível em: https://github.com/AnnaPaulaFigueiredo/avaliacao_llms_glosa/
[2] Disponível em: https://sol.sbc.org.br/index.php/sbsi/article/view/34361

Este trabalho, portanto, representa um ponto de partida para iniciativas mais amplas voltadas à construção de tecnologias linguísticas acessíveis. Os resultados obtidos demonstram o potencial do uso de modelos de linguagem para promover inclusão comunicacional, e abrem caminho para pesquisas futuras nas interseções entre linguística computacional, inteligência artificial e acessibilidade para a comunidade surda.

REFERÊNCIAS

ABDIN, M. *et al.* Phi-4 technical report. **arXiv preprint arXiv:2412.08905**, 2024. Acessado em junho de 2025. Disponível em: https://arxiv.org/pdf/2412.08905.

AI, M. **Meta Llama 3: The Next Generation of Open Source Large Language Models**. 2024. https://ai.meta.com/blog/meta-llama-3/. Acessado em março de 2025.

AL-GHURIBI, S. M.; NOAH, S. A. M. A Comprehensive Overview of Recommender System and Sentiment Analysis. **arXiv preprint arXiv:2109.08794**, 2021. Disponível em: https://arxiv.org/abs/2109.08794.

ALLEN, J. **Natural Language Understanding**. 1995. Benjamin Cummings Series in Computer Science. Disponível em: https://books.google.com.br/books?id=l0DOH4NHneIC. Acessado em: 08 de Janeiro de 2024.

ANDREAS, J.; VLACHOS, A.; CLARK, S. Semantic parsing as machine translation. **Association for Computational Linguistics**, Sofia, Bulgaria, p. 47–52, ago. 2013. Acessado em: janeiro de 2024. Disponível em: https://www.aclweb.org/anthology/P13-2009.

ARAUJO, M. d. S. O.; PEIXOTO, E. R. F.; SOUSA, R. d. S. N. d. **A Constituição da Identidade Linguística Surda: A Narrativização Avaliativa Surda Sobre Língua Portuguesa e Libras**. 2021. Repositório Institucional da UFPA. Disponível em: http://repositorio.ufpa.br/jspui/handle/2011/13611. Acessado em: 23 de Dezembro de 2023.

Associação dos Gerentes do Banco do Brasil. **Maioria dos bancos usa chatbot para atendimento no WhatsApp**. 2022. http://www.agebb.com.br/tag/chatbot/. Acessado em maio de 2024.

BAHAN, B. **Non-Manual Realization of Agreement in American Singn Language**. 1996. Tese de doutorado, Boston University. Disponível em: https://wayback.archive-it.org/all/20171011051748/ftp://louis-xiv.bu.edu/pub/asl/disserts/Bahan96.pdf. Acessado em 14 de Fevereiro de 2024.

BOOTH, W. C.; COLOMB, G. G.; WILLIAMS, J. M. **The Craft of Research**. Chicago: University of Chicago Press, 2008.

BOWMAN, S. R. Eight Things to Know about Large Language Models. **arXiv preprint arXiv:2304.00612**, 2023. Disponível em: https://arxiv.org/abs/2304.00612.

BRAGA, A. d. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. d. L. F. **Redes neurais artificiais: teoria e aplicações**. Rio de Janeiro: LTC, 2000.

BROWN, T. *et al.* Language models are few-shot learners. **Advances in Neural Information Processing Systems**, v. 33, p. 1877–1901, 2020.

CAIS, C. for A. S. **Massive Multitask Language Understanding (MMLU)**. 2023. Acessado em: Fev. 2025. Disponível em: https://huggingface.co/datasets/cais/mmlu.

CHAWRE, H. **Fine-Tuning LLMs: Overview, Methods, and Best Practices**. 2023. https://www.turing.com/resources/finetuning-large-language-models. Accessed: 2025-07-03.

CHEN, S. *et al.* **Real Learner Data Matters: Exploring the Design of LLM-Powered Question Generation for Deaf and Hard of Hearing Learners**. Honolulu, HI, USA: ACM, 2024.

CHEN, Y. *et al.* Textbooks are all you need ii: phi-1.5 technical report. **arXiv preprint arXiv:2306.11644**, 2023. Acessado em junho de 2025. Disponível em: https://arxiv.org/abs/2306.11644.

CHOUDHARY, P.; CHAUHAN, S. **An intelligent chatbot design and implementation model using long short-term memory with recurrent neural networks and attention mechanism**. 2023. 100359 p. Decision Analytics Journal. Disponível em: https://www.sciencedirect.com/science/article/pii/S2772662223001996. Acessado em fevereiro de 2024.

CONSTITUINTE, A. N. **Constituição da República Federativa do Brasil**. Brasília - DF: Diário Oficial da União, 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acessado em: dezembro de 2023.

CONTEúDO, E. **Itaú estuda como levar inteligência artificial generativa para atendimento ao cliente**. 2024. Acessado em: setembro 2024. Disponível em: https://www.moneytimes.com.br/itau-estuda-como-levar-inteligencia-artificial-generativa-para-atendimento-ao-cliente/.

COSTA, R. R. A. *et al.* **Acessibilidade na TV 3.0 Brasileira a partir de Mídias de Legenda, Glosa e Áudio Descrição**. Ribeirão Preto/SP: Sociedade Brasileira de Computação, 2023. 123-129 p. Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Disponível em: https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/25664. Acessado em: 22 de Dezembro de 2023.

DALE, R. The return of the chatbots. **Natural Language Engineering**, v. 22, p. 811–817, 09 2016.

DANIEL, H.; MICHAEL HAN; TEAM, U. **Unsloth**. 2023. Disponível em: https://github.com/unslothai/unsloth.

DEBEVC, M.; KOSEC, P.; HOLZINGER, A. Improving multimodal web accessibility for deaf people: Sign language interpreter module. **Multimedia Tools and Applications**, v. 54, p. 181–199, 08 2011.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **NAACL-HLT**, p. 4171–4186, 2019.

DIGITAL, G. **Acessibilidade Digital**. 2024. Brasília, DF: Governo Digital. Disponível em: https://www.gov.br/governodigital/pt-br/acessibilidade-digital. Acessado em janeiro de 2024.

Estatísticas Sociais. Pns 2019: país tem 17,3 milhões de pessoas com algum tipo de deficiência. **Agência IBGE**, 2021. Disponível em: https://encurtador.com.br/tBMT8. Acessado em: 17 de Dezembro de 2023.

FLICK, U. **Introducing Research Methodology: A Beginner's Guide to Doing a Research Project**. SAGE: Publications Limited, 2018.

GANDRA, A. **Lei que reconhece Libras como Língua Oficial do País completa 20 anos**. 2022. Agência Brasil. Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2022-04/lei-que-reconhece-libras-como-lingua-oficial-do-pais-completa-20-anos. Acessado em: 10 de Janeiro de 2024.

GAO, T.; FISCH, A.; CHEN, D. Making pre-trained language models better few-shot learners. 2021.

GOLDMAN, O. *et al.* Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. **arXiv preprint arXiv:2502.21228**, 2025. Disponível em: https://arxiv.org/abs/2502.21228.

GONCALVES, A.; FREIRE, A.; PEREIRA, D. **Accessibility Assessment in Banking Chatbots using Texts Written by Deaf People who use Portuguese as a Second Language**. 2025. 449-458 p.

GOODFELLOW, I. *et al.* Generative adversarial networks. **arXiv preprint arXiv:1406.2661**, 2014. Acessado em maio de 2025. Disponível em: https://arxiv.org/abs/1406.2661.

GOOGLE. **Todo mundo tem direito de acessar e aproveitar a Web**. 2024. Acessado em: fevereiro de 2024. Disponível em: https://www.google.com/intl/pt-BR/accessibility/.

GRANDA, A. **OMS estima 2,5 bilhões de pessoas com problemas auditivos em 2050**. 2021. Agência Brasil. Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2021-03/oms-estima-25-bilhoes-de-pessoas-com-problemas-auditivos-em-2050. Acessado em: 15 de Dezembro de 2023.

HADDON, L.; PAUL, G. **Technology and the Market: Demand, Users and Innovation**. 2001. 201-215 p. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/132/118. Acessado em: 22 de Dezembro de 2023.

HENDRYCKS, D. *et al.* Aligning ai with shared human values. **arXiv preprint arXiv:2008.02275**, 2021. Acessado em maio de 2024.

HENDRYCKS, D. *et al.* Measuring massive multitask language understanding. **arXiv preprint arXiv:2009.03300**, 2021.

HENDY, A. *et al.* How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. **arXiv preprint arXiv:2302.09210v1**, 2023. Disponível em: https://arxiv.org/abs/2302.09210v1.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, p. 328–339, 2018.

HU, E. J. *et al.* Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

HUSSAIN, S.; SIANAKI, O. A.; ABABNEH, N. A survey on conversational agents/chatbots classification and design techniques. **Springer International Publishing**, Cham, p. 946–956, 2019.

IBM. **Com BIA, Bradesco e IBM transformam o atendimento de milhões de usuários**. 2019. https://www.ibm.com/blogs/ibm-comunica/com-bia-bradesco-e-ibm-transformam-o-atendimento-de-milhoes-de-usuarios/. Acessado em maio de 2024.

International Organization for Standardization (ISO). **ISO 9241-171: Ergonomics of human-system interaction – Part 171: Guidance on software accessibility**. 2016. https://www.iso.org/standard/63500.html. Acessado em: 01 set. 2025.

JOKSIMOSKI, B. *et al.* **Technological Solutions for Sign Language Recognition: A Scoping Review of Research Trends, Challenges, and Opportunities**. 2022. 1-1 p. IEEE Access. Disponível em: https://ieeexplore.ieee.org/document/9569151. Acessado em: 05 de Janeiro de 2024.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. New York: Pearson, 2019. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. Acessado em: 17 de Dezembro de 2023. ISBN 978-0131873216.

KHENNOUCHE, F. *et al.* Revolutionizing customer interactions: Insights and challenges in deploying chatgpt and generative chatbots for faqs. **arXiv preprint arXiv:2311.09976**, 2023. Disponível em: https://arxiv.org/abs/2311.09976.

LATENODE. **Zephyr-7B: From Transformer Architecture to Practical Implementation**. 2024. https://latenode.com/blog/zephyr-7b-from-transformer-architecture-to-practical-implementation. Accessed: 2025-06-29.

LEITE, T. d. A. *et al.* **Semântica lexical na libras: Libertando-se da tirania das glosas**. 2022. 1-23 p. Revista da ABRALIN. Disponível em: https://revista.abralin.org/index.php/abralin/article/view/1833. DOI: DOI: 10.25189/rabralin.v20i3.1833.

LIBRARIA. **Comunicação Inclusiva: Uma Necessidade Vital no Mundo de Hoje**. 2023. Acessado em: 08 de Fevereiro de 2024. Disponível em: https://libraria.com.br/comunicacao-inclusiva/.

LIDDY, E. D. **Natural Language Processing**. New York: Marcel Decker, Inc., 2001.

MAGLOGIANNIS, I.; ILIADIS, L.; PIMENIDIS, E. **An Overview of Chatbot Technology**. 2020. 373–383 p. National Library of Medicine. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256567/. Acessado em: dezembro de 2023.

MANNING, C. D.; SCHüTZE, H. **Foundations of Statistical Natural Language Processing**. 1999. MIT Press.

MARKOWITZ, D.; WARKENTIN, T.; DYER, E. **Prompt Engineering Patterns: Best Practices for Using Language Models**. 2023. Disponível em: https://developers.googleblog.com/2023/03/prompt-engineering-best-practices.html.

MCCLEARY, L. E.; VIOTTI, E. d. C. **Transcrição de dados de uma língua sinalizada: Um estudo piloto da transcrição de narrativas na língua de sinais brasileira (LSB)**. 2007. Cânome Editorial. Disponível em: https://repositorio.usp.br/item/001697317. Acessado em: 26 de Dezembro de 2023.

MICROSOFT. **Nossa abordagem de acessibilidade**. 2024. Acessado em: fevereiro de 2024. Disponível em: https://www.microsoft.com/pt-br/accessibility/approach.

MIKOLOV, T. *et al.* Distributed representations of words and phrases and their compositionality. **arXiv preprint arXiv:1310.4546**, 2013. Acessado em fevereiro de 2024. Disponível em: https://arxiv.org/abs/1310.4546.

MORAES, R. G. C. d. **Agentes Conversacionais Inteligentes e a Comunicação Mediada por Algoritmos: As Complexidades das Interações com os Usuários no Cotidiano**. Tese (Doutorado) — Pontifícia Universidade Católica de São Paulo, 2023. Disponível em: https://repositorio.pucsp.br/jspui/handle/handle/32544.

MOREIRA, A.; PAZOTI, M. A.; SISCOUTTO, R. A. Easy asistant: uma ferramenta de auxílio para o desenvolvimento de chatbots de domínimo específico. **Colloquium Exactarum. ISSN: 2178-8332**, v. 13., n. 3., p. 48–58, fev. 2022. Disponível em: https://revistas.unoeste.br/index.php/ce/article/view/4105.

MULDOWNEY, O. **CHATBOTS: An Introduction And Easy Guide To Making Your Own**. Dublin, Ireland: Curses & Magic, 2017. First published 2017 by Curses & Magic, Dublin, Ireland. ISBN 978-1-9998348-0-7.

NASCIMENTO, L. R. S.; COSTA, E. d. S. **Escrita de sinais no Brasil e suas interfaces**. Porto Velho, RO: Edufro – Editora da Universidade Federal de Rondônia, 2022.

NASCIMENTO, M. V. B. **Formação de intérpretes de Libras e Língua Portuguesa: encontros de sujeitos, discursos e saberes**. Tese (Doutorado) — Pontifícia Universidade Católica de São Paulo, São Paulo, 2016. Acessado em 18 de agosto de 2023. Disponível em: https://tede2.pucsp.br/handle/handle/19562.

NASCIMENTO, V. Do campo ao texto em pesquisas com tradutores e intérpretes de libras-português: desafios na transcrição de corpora bilíngue intermodal simultâneo. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, Pontifícia Universidade Católica de São Paulo - PUC-SP, v. 39, n. 4, p. 202339458715, 2023. ISSN 0102-4450. Disponível em: https://doi.org/10.1590/1678-460X202339458715.

NUNES, I. F.; LACERDA, N. A. Novos letramentos e inclusão de alunos surdos: o chatgpt no contexto de tecnologia assistiva para o ensino de português como l2 na modalidade escrita. **Dialogia**, Universidade Paulista - UNIP, v. 52, n. 52, p. 1–21, 2025. ISSN 1983-9294.

OLIVEIRA, E. S. de; LIMA, M. A. C. B.; ARAUJO, T. M. U. de. Uma estratégia de tradução automática de textos em língua portuguesa para glosa em língua brasileira de sinais. **Sociedade Brasileira de Computação**, Manaus, p. 23–26, 2015. ISSN 2596-1683.

OTHMAN, A.; JEMNI, M. Statistical sign language machine translation: from english written text to american sign language gloss. **arXiv preprint arXiv:1112.0168**, 2011.

PASSERO, G.; ENGSTER, N. E. W.; DAZZI, R. L. S. Uma revisão sobre o uso das tics na educação da geração z. **Revista Novas Tecnologias na Educação**, v. 14., n. 2, dez. 2016. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/70652.

PENG, B. *et al.* RWKV: Reinventing RNNs for the Transformer Era. **arXiv preprint arXiv: 2305.13048**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2305.13048.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. **GloVe: Global Vectors for Word Representation**. Doha, Qatar: ACL, 2014. 1532–1543 p. Disponível em: https://aclanthology.org/D14-1162.

PETROV, A. *et al.* Language Model Tokenizers Introduce Unfairness Between Languages. **arXiv preprint arXiv:2305.15425**, 2023. Disponível em: https://arxiv.org/abs/2305.15425.

PFEIFFER, J. *et al.* Adapterfusion: Non-destructive task composition for transfer learning. **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, p. 1246–1257, 2020.

PHAM, V. *et al.* Dropout improves Recurrent Neural Networks for Handwriting Recognition. **arXiv preprint arXiv: 1312.4569**, 2013. Disponível em: https://arxiv.org/abs/1312.4569.

QUADROS, R. M. d.; SOARES, J. S. d.; MIRANDA, R. D. Id-sinais para organização e busca de corpus em libras. In: STUMPF, M.; LEITE, T. d. A.; QUADROS, R. M. d. (Ed.). **Estudos da Língua Brasileira de Sinais**. Florianópolis: Insular, 2014. v. 2, p. 29–44. Disponível em: https://shorturl.at/fsyGW. Acesso em 28 de fevereiro de 2021.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. **OpenAI Blog**, 2018. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. **OpenAI Blog**, 2018. Acessado em: Fev. 2025. Disponível em: https://openai.com/blog/language-unsupervised.

RAFAILOV, R. *et al.* Direct preference optimization: Your language model is secretly a reward model. **arXiv preprint arXiv:2212.09526**, 2023. Disponível em: https://arxiv.org/abs/2212.09526.

RAFFEL, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020.

REPúBLICA, P. da. **Lei Brasileira de Inclusão das Pessoas com Deficiência**. Brasília, 2015. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm. Acessado em: dezembro de 2023.

RODRIGUES, C. H. Competência em tradução e línguas de sinais: a modalidade gestual-visual e suas implicações para uma possível competência tradutória intermodal. **Trabalhos em Linguística Aplicada**, v. 57, p. 287–318, Março 2018. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/tla/article/view/8651578. Acessado em: 20 de Dezembro de 2023.

SACHDEVA, N. *et al.* How to train data-efficient llms. **arXiv preprint arXiv:2402.09668**, 2024. Disponível em: https://arxiv.org/pdf/2402.09668.

SARDENBERG, L. F.; BUOGO, S. **Cerca de um Bilhão de Pessoas com Deficiência têm Acesso Negado Tecnologia Assistiva**. 2022. Nações Unidas do Brasil. Disponível em: https://brasil.un.org/pt-br/182189-cerca-de-um-bilh%C3%A3o-de-pessoas-com-defici%C3%AAncia-t%C3%AAm-acesso-negado-tecnologia-assistiva. Acessado em: 15 de Dezembro de 2023.

SCIMAGO INSTITUTIONS RANKINGS. Inteligência artificial: riscos, benefícios e uso responsável. **Estudos Avançados**, SciELO Brazil, v. 35, n. 101, p. 187–200, 2021. Disponível em: https://www.scielo.br/j/ea/a/ZnKyrcrLVqzhZbXGgXTwDtn/.

SHAHIN, N.; ISMAIL, L. Chatgpt, let us chat sign language: Experiments, architectural elements, challenges and research directions. **arXiv preprint arXiv:2401.06804**, 2024. Disponível em: https://arxiv.org/abs/2401.06804.

SLOBIN, D. I. **Quebrando modelos: As Línguas de Sinais e a Natureza da Linguagem Humana**. 2015. 844-853. p. Disponível em: https://periodicos.ufsc.br/index.php/forum/article/view/1984-8412.2015v12n3p844.. Acessado em: 20 de Dezembro de 2023.

SOLEMAN, C.; BOUSQUAT, A. Políticas de saúde e concepções de surdez e de deficiência auditiva no sus: um monólogo? **Cadernos de Saúde Pública**, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 37, n. 8, p. e00206620, 2021. ISSN 0102-311X. Disponível em: https://doi.org/10.1590/0102-311X00206620.

TALEB, I. *et al.* Big data quality framework: A holistic approach to continuous quality management. **Journal of Big Data**, v. 8, n. 76, 2021. Disponível em: https://www.aeaweb.org/articles?id=10.1257/aer.91.4.909. Acessado em: janeiro de 2023.

TAORI, R. *et al.* Alpaca: A strong, replicable instruction-following model. **Stanford CRFM**, 2023. Disponível em: https://crfm.stanford.edu/2023/03/13/alpaca.html.

TEAM, Q. Qwen technical report. **arXiv preprint arXiv:2309.16609**, 2023. Disponível em: https://arxiv.org/abs/2309.16609.

TEAM, Q. Qwen2: The next generation of qwen models. **arXiv preprint arXiv:2412.15115**, 2024. Acessado em março de 2025. Disponível em: https://arxiv.org/pdf/2412.15115.

TEAM, Q. Qwen3: High-quality, open multilingual language models and instruction-tuned models. **arXiv preprint arXiv:2505.09388**, 2025. Acessado em junho de 2025. Disponível em: https://arxiv.org/pdf/2505.09388.

TOUVRON, H. *et al.* Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023. Disponível em: https://arxiv.org/abs/2302.13971.

TUNSTALL, L. *et al.* Zephyr: Direct distillation of lm alignment. **arXiv preprint arXiv:2310.16944**, 2023. Disponível em: https://arxiv.org/abs/2310.16944.

TURING, A. M. Computing machinery and intelligence. **Mind**, v. 49, p. 433–460, 1950.

VASWANI, A. *et al.* Attention is all you need. **arXiv preprint arXiv:1706.03762**, p. 5998–6008, 2017. Acessado em abril de 2025. Disponível em: https://arxiv.org/abs/1706.03762.

VASWANI, A. *et al.* Attention is all you need. **arXiv preprint arXiv:1706.03762v7**, 2021. Acessado em: março 2025. Disponível em: https://arxiv.org/abs/1706.03762v7.

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in NLP. **Aclanthology**, 1992. Acessado em: março de 2024. Disponível em: https://aclanthology.org/C92-4173.

WOLF, T. *et al.* Transformers: State-of-the-art natural language processing. Association for Computational Linguistics, p. 38–45, 2020. Disponível em: https://aclanthology.org/2020.emnlp-demos.6.

World Wide Web Consortium (W3C). **Web Content Accessibility Guidelines (WCAG) 2.2**. 2023. https://www.w3.org/TR/WCAG22/. Acessado em: 01 set. 2025.

XU, C. *et al.* Wizardlm: Empowering large language models to follow complex instructions. **arXiv preprint arXiv:2304.12244**, 2023. Acessado em julho de 2024. Disponível em: https://arxiv.org/abs/2304.12244.

ZAREMBA, W.; SUTSKEVER, I. Learning to generate reviews and discover sentiment. **arXiv preprint arXiv:1406.1827**, 2014. Acessado em junho de 2024.

ZEMČÍK, M. T. A brief history of chatbots. **DEStech Transactions on Computer Science and Engineering**, DEStech Publishing Inc. Lancaster, PA, USA, v. 10, 2019.

ZHU, H. Architectural foundations for the large language model infrastructures. **arXiv preprint arXiv:2408.09205**, 2024. Disponível em: https://arxiv.org/pdf/2408.09205v2.