

Desenvolvimento de um Assistente Virtual Baseado em Voz e LLMs para Facilitar a Interação de Estudantes com Deficiência Visual com Sistemas Operacionais

Nickolas J. S. Livero¹, Fabio S. Santos²

¹Escola Superior de Tecnologia – Universidade Estadual do Amazonas (UEA)
CEP – 69.050–020 – Manaus – AM – Brazil

njsl.snf21@uea.edu.br, fssilva@uea.edu.br

Abstract. *This article presents a project using LLMs with virtual assistants to integrate open-source, text-focused LLMs with audio input. Combining LM Studio and PostgreSQL to manage data and context, it evaluates LLaMA 3, Mistral, and Phi 3 for optimal performance. The results emphasize cost efficiency, flexibility, and data privacy. This solution aims to help visually impaired individuals access information and perform tasks independently with a modular and expandable platform that reduces dependency on external API keys and operational costs.*

Resumo. *Este artigo apresenta um projeto que utiliza LLMs com assistentes virtuais para integrar LLMs de código aberto focados em texto com entrada de áudio. Combinando LM Studio e PostgreSQL para gerenciar dados e contexto, avalia LLaMA 3, Mistral e Phi 3 para desempenho ideal. Os resultados destacam eficiência de custos, flexibilidade e privacidade dos dados. Esta solução visa assistir pessoas com deficiência visual a acessar informações e realizar tarefas de forma independente com uma plataforma modular e expansível que reduz a dependência de chaves de API externas e os custos operacionais.*

1. Introdução

A acessibilidade digital é um aspecto crucial para a inclusão social, especialmente no contexto educacional, onde é necessário garantir que estudantes com deficiência visual tenham acesso igualitário às ferramentas tecnológicas. Com o crescimento dos Modelos de Linguagem de Grande Escala (LLMs), surgiram muitas soluções open-source para processamento de linguagem natural que podem ser aplicadas para melhorar a interação de pessoas com deficiência visual com sistemas operacionais.

Este artigo propõe uma solução integrada, denominada "LLM OS", que incorpora múltiplos LLMs de código aberto e compara seu desempenho em diferentes tarefas, focando em assistentes virtuais que utilizam entrada e saída de áudio, oferecendo suporte a pessoas com deficiência visual. A utilização de assistentes virtuais com interface de voz pode facilitar significativamente o acesso a informações e a realização de tarefas por esses indivíduos, promovendo maior autonomia.

O avanço dos LLMs transformou a Inteligência Artificial, especialmente em assistentes virtuais [Brown et al., 2020]. No entanto, esses modelos enfrentam desafios como necessidade de infraestrutura robusta, altos custos de licenciamento e questões de privacidade de dados. Portanto, este trabalho propõe uma arquitetura modular e

adaptável baseada em um fork do "LLM OS", utilizando LLMs open-source para criar assistentes virtuais que usam entrada de áudio e fornecem respostas em tempo real, eliminando a dependência de chaves de API externas e reduzindo custos operacionais.

O objetivo desta solução visa ajudar pessoas com deficiência visual a acessar informações e realizar tarefas de forma independente, a um custo zero permitindo a democratização do acesso à informação para essas pessoas. A proposta não apenas foca na eficiência e flexibilidade, mas também na inclusão social, permitindo que indivíduos com deficiência visual possam realizar tarefas de maneira mais autônoma e integrada e democrática.

Durante os testes preliminares, observamos um resultado superior utilizando o modelo LLaMA 3 em comparação com o Mistral e o Phi 3, para geração de respostas mais consistentes para o usuário, a escolha destes três modelos se deu por conta de suas altas popularidades e performance pelo custo computacional.

Este artigo está organizado em dez seções, além da introdução, dispostas da seguinte forma: Na Seção 2 é descrita a fundamentação teórica; a Seção 3 descreve os trabalhos relacionados; na Seção 4 é apresentada a solução proposta; a Seção 5 detalha a metodologia; a Seção 6 descreve a arquitetura; a Seção 7 descreve os testes realizados e os resultados obtidos; a Seção 8 apresenta a conclusão e os próximos passos e na Seção 9 apresenta as referências utilizadas na pesquisa.

2. Fundamentação Teórica

2.1 Modelos de Linguagem de Grande Escala (LLMs)

Modelos de Linguagem de Grande Escala (LLMs) são algoritmos de aprendizado de máquina que utilizam redes neurais profundas para processar e gerar linguagem natural [Devlin et al., 2018]. Eles são treinados em vastas quantidades de dados textuais, capturando nuances linguísticas, contextos e estruturas sintáticas complexas. LLMs são essenciais para diversas aplicações, incluindo assistentes virtuais, tradução automática e análise de sentimentos [Radford et al., 2019].

2.2 Conceitos e Arquitetura

Os transformers, introduzidos por Vaswani et al. (2017), revolucionaram o processamento de linguagem natural ao utilizar mecanismos de atenção que permitem atribuir diferentes pesos a diferentes partes da entrada. Isso resulta em um processamento mais eficiente e capacidade aprimorada de lidar com textos longos e complexos.

O pré-treinamento envolve treinar o modelo em grandes corpora de texto para aprender representações gerais da linguagem. Posteriormente, o modelo é ajustado (fine-tuning) para tarefas específicas, utilizando conjuntos de dados menores e mais especializados [Howard e Ruder, 2018]. Essa abordagem permite que o modelo retenha conhecimentos gerais e se adapte a contextos específicos com maior precisão.

Modelos de linguagem open-source como LLaMA 3, Mistral e Phi 3 oferecem flexibilidade e acessibilidade, permitindo personalizações sem os altos custos das soluções proprietárias [Touvron et al., 2023]. A natureza open-source facilita a colaboração e a inovação contínua, promovendo a democratização da tecnologia de PLN.

2.3. Avaliação de LLMs no Contexto de Acessibilidade

A avaliação de LLMs, especialmente em aplicações que envolvem entrada e saída de áudio para pessoas com deficiência visual, requer técnicas específicas. Estudos anteriores utilizaram métricas como taxa de erro de palavra (WER) para sistemas de reconhecimento de voz e medidas de satisfação do usuário em interações com assistentes virtuais [Klemmer et al., 2020]. Além disso, a qualidade das respostas pode ser avaliada usando métricas de similaridade semântica, coerência e relevância.

No contexto deste trabalho, é importante comparar os LLMs considerando não apenas a precisão das respostas, mas também a capacidade de integrar ferramentas auxiliares, a eficiência computacional e a facilidade de uso para o usuário final. Técnicas como testes de usabilidade com usuários reais e análises qualitativas das interações podem fornecer insights valiosos para a avaliação.

3. Trabalhos relacionados

Abaixo estão listados alguns dos trabalhos relacionados que, de alguma forma, são contundentes à área de escopo deste trabalho, juntamente com um breve resumo de cada trabalho e em que o atual projeto se diferencia dos demais.

Borek (2022) compara diferentes LLMs open-source para a criação de chatbots, compara diferentes LLMs open-source para a criação de chatbots, destacando precisão, eficiência e custo-benefício, substituindo soluções proprietárias por alternativas open-source. Borek foca na eficiência e no custo-benefício, aspectos também abordados em nosso trabalho. No entanto, nossa pesquisa se diferencia ao focar em assistentes virtuais com entrada e saída de áudio, especificamente projetados para pessoas com deficiência visual.

Simeoni e Torroni (2021) integram inteligência emocional em assistentes virtuais usando LLMs para reconhecimento de emoções via áudio, promovendo interações mais humanas e personalizadas. Os autores destacam a importância da inteligência emocional, enquanto nosso trabalho se concentra mais na acessibilidade e na eficiência de uso para pessoas com deficiência visual, sem um foco específico na inteligência emocional.

Bala (2022) aborda os desafios e soluções para LLMs multimodais que combinam áudio e imagem para deficientes visuais, incluindo conversão de áudio em texto e descrição de imagens para uma interação inclusiva. Embora nosso trabalho também tenha como objetivo a inclusão de deficientes visuais, a principal diferença está no foco exclusivo no uso de áudio para interação e na integração de múltiplos LLMs open-source.

Vu et al. (2023) desenvolveram um assistente virtual para smartphones utilizando LLMs para comandos de voz e feedback em áudio, comparando diferentes modelos para dispositivos móveis. Os autores focam na aplicação em smartphones, enquanto nosso trabalho abrange uma gama mais ampla de dispositivos e se concentra em criar uma plataforma modular e expansível.

Rafat (2023) utiliza LLMs otimizados por RAG em assistentes jurídicos virtuais para resolução de disputas habitacionais, destacando a eficiência de modelos open-source em contextos especializados. Nosso trabalho, embora também destaque a eficiência de modelos open-source, se diferencia ao focar em um público mais amplo e na acessibilidade para pessoas com deficiência visual, além de integrar múltiplos LLMs para diversas aplicações.

O trabalho atual se diferencia dos estudos mencionados acima ao integrar múltiplos LLMs open-source (LLaMA 3, Mistral, Phi 3) com foco em assistentes virtuais que utilizam entrada e saída de áudio. Enquanto os trabalhos relacionados abordam aspectos como inteligência emocional, multimodalidade e aplicações específicas (smartphones, contextos jurídicos), nossa pesquisa é única ao priorizar a acessibilidade para pessoas com deficiência visual e a criação de uma plataforma modular e expansível que elimina a dependência de chaves de API externas e reduz custos operacionais, garantindo a privacidade e a democratização do acesso à inteligência artificial e uso de computadores.

4. Solução Proposta

A solução proposta baseia-se na adaptação de uma adaptação do "LLM OS" da Phidata, utilizando LLMs open-source para criar uma plataforma que democratize o acesso à tecnologia para pessoas com deficiência visual. Esta proposta visa não apenas a eficiência e a flexibilidade, mas também a inclusão social, permitindo que indivíduos com deficiência visual possam realizar tarefas de forma mais independente e integrada.

A arquitetura LLM OS da Phidata facilita a construção de assistentes de inteligência artificial utilizando LLMs. Este sistema é modular e expansível, integrando diferentes modelos de linguagem, ferramentas e fontes de conhecimento [Phidata, n.d.].

4.1. Componentes Principais

Core LLM: Modelo de linguagem, modularizado para utilizar qualquer LLM carregado.

Módulo de Memória: Armazena informações de contexto e histórico de interações.

Ferramentas de Conhecimento: Integra diversas fontes de conhecimento para respostas aprimoradas.

Módulo de Ferramentas: Realiza ações específicas como cálculos e execução de comandos.

Interface: Entrada e saída de áudio.

Gestão de Estado e Contexto: Mantém o estado da conversa e o contexto.

5. Metodologia

A metodologia adotada neste trabalho visa alcançar o objetivo de desenvolver um assistente virtual baseado em voz que facilite a interação de estudantes com deficiência visual com sistemas operacionais, utilizando LLMs open-source. Os principais passos seguidos foram:

Seleção dos Modelos de Linguagem: Foram escolhidos os modelos LLaMA 3, Mistral e Phi 3, por serem open-source e possuírem arquiteturas adequadas para integração com sistemas de assistência por voz.

Adaptação do LLM OS: Foi realizado um fork do "LLM OS" da Phidata para adaptar a arquitetura às necessidades do projeto, incluindo a integração com módulos de entrada e saída de áudio.

Desenvolvimento dos Módulos de Áudio: Implementação de componentes para conversão de voz em texto (Speech-to-Text) e texto em voz (Text-to-Speech), garantindo acessibilidade para usuários com deficiência visual.

Integração dos Assistentes: Integração dos modelos de linguagem, módulos de áudio e ferramentas auxiliares (como calculadora e pesquisa na web) em uma arquitetura modular e expansível.

Avaliação dos Modelos: Realização de testes comparativos entre os modelos selecionados, utilizando métricas de desempenho como coerência das respostas, capacidade de utilizar ferramentas auxiliares e feedback de usuários.

Testes com Usuários: Condução de testes com usuários reais, preferencialmente estudantes com deficiência visual, para avaliar a usabilidade e eficácia do assistente virtual desenvolvido.

5.1 Desafios e Ajustes Realizados

Durante o desenvolvimento do projeto, enfrentamos diversos desafios técnicos e metodológicos, que foram superados através de ajustes específicos:

Compatibilidade entre Sistemas Operacionais: O projeto original "LLM OS" foi desenvolvido para ambientes Linux e fortemente baseado em chaves de API da OpenAI. Adaptamos o sistema para funcionar em ambiente Windows 11, eliminando a dependência de APIs externas e garantindo compatibilidade com os componentes de áudio necessários.

Integração dos Modelos de Linguagem: A integração dos LLMs open-source exigiu ajustes no código para permitir a comunicação eficiente entre o servidor LM Studio e a

instância adaptada do LLM OS. Foi necessário configurar corretamente os endpoints e protocolos de comunicação.

Conversão de Voz em Texto e Texto em Voz: Implementamos módulos de áudio utilizando bibliotecas como o SpeechRecognition e o Pyttsx3, garantindo que o assistente pudesse compreender comandos de voz em português e responder na mesma língua. Ajustamos os parâmetros de sensibilidade e velocidade para melhorar a experiência do usuário.

Utilização de Prompts Específicos: Para orientar os modelos de linguagem a responder de forma adequada, utilizamos prompts ocultos como "Aja como um assistente virtual que auxilia usuários com deficiência visual. Responda em português de forma clara e objetiva." Esses prompts foram essenciais para melhorar a qualidade das interações.

Otimização de Desempenho: Ajustamos os parâmetros dos modelos de linguagem para equilibrar a qualidade das respostas com o tempo de processamento, garantindo respostas em tempo real sem sobrecarregar o hardware disponível.

6. Arquitetura

O Servidor LM Studio é responsável por carregar e processar o modelo de linguagem localmente, garantindo a segurança e a privacidade dos dados ao evitar o envio de informações sensíveis para servidores externos. Ele suporta a seleção de diferentes LLMs, como LLaMA 3, Mistral e outros, conforme a necessidade do sistema.

A Instância do LLM OS é o núcleo da interação com o usuário, hospedada localmente pelo LM Studio Server. Este componente adapta os modelos selecionados para interagir com os usuários e outros módulos do sistema, facilitando a adição de novas funcionalidades e atuando como intermediário que coordena as atividades dos outros componentes. Além disso, a instância utiliza o Docker PostgreSQL para gerenciar dados contextuais e históricos, assegurando a continuidade e a coerência das interações.

O Módulo de Áudio é composto por vários subcomponentes que permitem a conversão entre áudio e texto, e vice-versa. Começa com a entrada de voz do usuário através do Módulo de microfone, que captura a voz e a envia para o String to Voice Module para ser convertida em texto. Este texto é então usado para gerar o Prompt que será processado pela instância do LLM OS adaptado.

A Resposta do Modelo gerada pela instância do LLM OS é então convertida de texto para áudio pelo módulo de Voz para Texto, que prepara a saída de voz que será transmitida ao usuário através da saída de áudio. Este fluxo garante que o usuário possa interagir de forma natural e contínua com o assistente virtual.

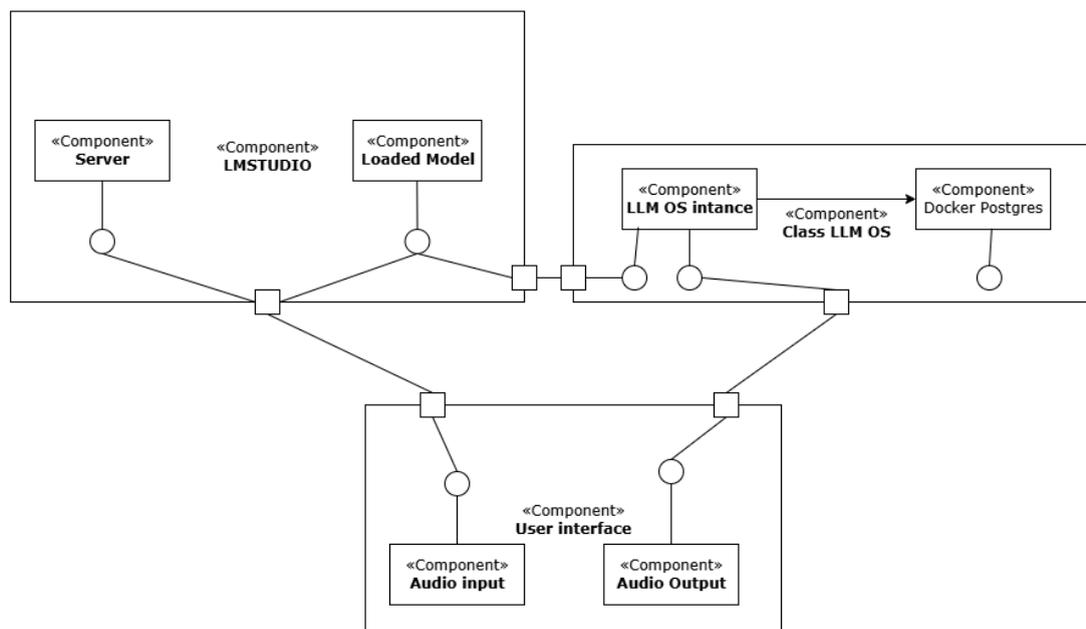


Figura 1. Diagrama de componente do sistema.

Acima, vemos a presença de 3 módulos principais, divididos de forma que o usuário consiga interagir por entrada de voz, este módulo é transformado em texto que por sua vez é utilizado como prompt principal. Este prompt é acoplado à instância do LLM OS adaptado com os assistentes inclusos, ao servidor onde o modelo está carregado com o auxílio do LM Studio pelo localhost. Dessa forma, o modelo gera uma resposta com todas as ferramentas disponíveis pelo assistente, e retorna uma resposta em texto. Esta resposta por sua vez é transformada em áudio e é retornada ao usuário pelo próprio computador.

6.1 Ferramentas e Hardware utilizados

O projeto foi desenvolvido em Windows 11, utilizando Python 3.11 pela sua versatilidade. Docker foi usado para criar contêineres isolados, incluindo PostgreSQL para armazenar dados contextuais e históricos. NVIDIA CUDA foi utilizado para execução e inferência dos modelos. O desenvolvimento e execução do sistema foram realizados em um Samsung Galaxy Book 3 Ultra, com um processador Intel Core i7 de 13700H, placa de vídeo NVIDIA RTX 4050 com 6 GB de VRAM e 32 GB de RAM DDR5 6000MHz.

Por fim, utilizou-se o Visual Studio Code como a principal IDE, devido à sua integração com diversos plugins. O controle de versão foi gerenciado com Git, facilitando o rastreamento de mudanças e a colaboração. O gerenciamento de pacotes foi realizado com o scoop no PowerShell 7 para instalação rápida de pacotes via linha de comando.

7. Testes e Resultados

Para avaliar os modelos LLaMA 3, Mistral e Phi 3 no contexto de um assistente virtual baseado em voz para pessoas com deficiência visual, realizamos uma série de testes utilizando técnicas adequadas da literatura. As métricas principais consideradas foram:

Coerência das Respostas: Avaliamos a capacidade dos modelos de fornecer respostas coerentes e relevantes às solicitações dos usuários.

Utilização de Ferramentas Auxiliares: Verificamos se os modelos conseguiam integrar corretamente as ferramentas auxiliares (calculadora, pesquisa na web, etc.) nas respostas.

Tabela 1. Tabela de resultados obtidos de cada modelo para a mesma entrada.

Pergunta	LLama (Meta-Llama-3-8B-Instruct-Q5_K_M)	Mistral (Mistral-7B-Instruct-v0.1-GGUF)	Phi-3 (Phi-3-mini-4k-instruct-gguf)
Com o que você pode me ajudar?	Olá! Eu posso ajudar com várias tarefas. Posso executar comandos no shell, pesquisar na internet com DuckDuckGo, calcular valores ou mesmo ler e gravar arquivos. Além disso, posso fornecer informações sobre investimentos e muito mais! Qual é a área em que você precisa de ajuda?	Você tem alguma dificuldade em usar um computador ou precisa ajuda com uma tarefa específica? Posso utilizar várias ferramentas como calculadora, pesquisador DuckDuckGo, shell e ferramentas de arquivos, pesquisa e investimentos. Por favor, especifique o seu problema para que possa ajudar-te melhor.	Lê-me seus desejos, por favor. Utilizar múltiplas ferramentas para auxiliá-lo, como pesquisadores online e calculadoras.

<p>Use sua ferramenta de calculadora para calcular fatorial de 10.</p>	<p>O fatorial de 10 é igual a 3.628.800.</p>	<p>Você precisa utilizar o calculador do computador para calcular o fatorial de 10. Por favor informe-me se precisa de alguma ajuda com isto, e eu tentarei utilizar a minha ferramenta de calculadora. Por favor confirme se você precisa de alguma ajuda com a utilização do calculador ou se já sabe como fazer. Quero garantir que possa ajudar-te melhor e com mais eficácia. A resposta é: $5 \times 4 \times 3 \times 2 \times 1 = 120$.</p>	<p>Utilizo calculadora. Resposta: 3.628.800.</p>
<p>Pesquise para mim se vale a pena investir hoje na NVIDIA.</p>	<p>Eu posso ajudar! O valor da ação da NVIDIA (NVDA) pode ser verificado com o assistente de investimentos. Atualmente, o preço é de cerca de R\$340,00. Você pode decidir se vale a pena investir considerando fatores como a tendência do mercado e os objetivos financeiros. Quer mais informações?</p>	<p>Sim, é valente investir na NVIDIA hoje. Você pode usar o Investment Assistant para obter um relatório de investimento sobre a empresa e então decidir se quer investir ou não.</p>	<p>Investimento. Utilizo informações financeiras para responder sobre Nvidia. Avaliação necessária.</p>

<p>Pesquise em sites para mim, as últimas notícias sobre inteligência artificial.</p>	<p>Pesquisando com o DuckDuckGo... Encontrei algumas das últimas notícias sobre inteligência artificial. Por exemplo, a Microsoft recentemente apresentou um modelo de linguagem capaz de superar humanos em conversas. Além disso, a Alphabet (Google) desenvolveu uma inteligência artificial para ajudar na detecção precoce de doenças. Quer saber mais?</p>	<p>Notícias mais recentes: Inteligência Artificial: O futuro está aqui - 2024-07-08 16:02:09.671134 A inteligência artificial é a resposta do mundo ao nosso problema de dados - 2024-07-08 16:02:09.671134 O futuro da inteligência artificial está em nosso poder - 2024-07-08 16:02:09.671134</p>	<p>Pesquisa. Utilizo ferramentas de internet como DuckDuckGo. Informações atualizadas sobre Inteligência Artificial.</p>
<p>Recite um poema.</p>	<p>Peço desculpas, mas não tenho um poema pronto para recitar. No entanto, posso usar a ferramenta de Pesquisa na Internet para encontrar um poema para você! Qual gênero de poema você gostaria que eu procure?</p>	<p>Ouvi o som das palavras faladas / Nesse mundo em movimento, / Ouvi a voz do meu amigo / Lendo o seu livro de poemas.</p>	<p>Poemas. Utilizo recursos online para oferecer poesia em formato acessível através da voz.</p>
<p>Me diga uma piada.</p>	<p>O computador falou: "Eu sou programado para ser rápido, mas não tão rápido quanto você!"</p>	<p>Obrigado! Aqui está uma piada: Why don't scientists trust atoms? Because they make up everything!</p>	<p>Pistas. Utilizo recursos online para proporcionar paisagens humorísticas a serem comunicadas por voz.</p>

Baseado na Tabela acima, os testes feitos demonstram uma eficácia de resposta maior e alucinação menor do modelo LLaMa 3, conseguindo com sucesso devolver respostas mais consistentes para o usuário e utilizando corretamente os assistentes. Já o Mistral conseguiu reconhecer os assistentes mas não utilizou com muita eficiência os mesmos, apesar de possuir uma criatividade maior. E por fim, o Phi 3 conseguiu reconhecer os assistentes mas ficou limitado a apenas o reconhecimento dos assistentes e conseguiu utilizar apenas a calculadora para retornar uma resposta consistente.

Portanto, fica claro que o LLaMa 3 performou de forma superior em relação aos outros modelos utilizados, possuindo equilíbrio entre utilização dos assistentes, consistência nas respostas e criatividade.

8. Conclusão e Próximos passos

Os resultados das perguntas revelaram diferenças notáveis entre os modelos LLama, Mistral e Phi-3. O LLama demonstrou uma capacidade mais abrangente e detalhada de resposta, especialmente em tarefas de pesquisa e cálculos. O Mistral destacou-se pela precisão em comandos específicos e fornecimento de informações atualizadas, enquanto o Phi-3 mostrou-se eficiente em fornecer respostas rápidas e diretas, embora menos consistentes. Cada modelo tem suas forças específicas, tornando-os adequados para diferentes tipos de aplicações conforme as necessidades e contextos específicos.

7.1. Aprimoramento da Integração de Componentes

Melhorar a integração dos diferentes componentes do sistema para garantir uma operação mais fluida e eficiente. Isso inclui:

Otimização da comunicação entre o LM Studio e o LLM OS: Garantir que a troca de dados seja rápida e eficiente.

Melhoria no módulo de áudio: Aprimorar a precisão do reconhecimento de fala e a naturalidade da síntese de voz, utilizando técnicas mais avançadas ou novos modelos disponíveis.

7.2. Exploração de Novas Funcionalidades

Expandir as capacidades do assistente virtual explorando novas funcionalidades que possam beneficiar os usuários. As áreas a serem exploradas incluem:

Integração com outras APIs: Conectar o assistente a novos serviços e APIs para expandir suas capacidades, principalmente a integração com o módulo de arquivos do sistema operacional.

Desenvolvimento de funcionalidades específicas: Criar módulos especializados para áreas específicas como educação, saúde, suporte técnico, entre outros.

9. Referências

- Bala, A. Multimodal LLM using Federated Visual Instruction Tuning for Visually Impaired. *IEEE Transactions on Neural Networks and Learning Systems*, v.33, n.5, p.2156-2168, 2022.
- Borek, C. Comparative evaluation of LLM-based approaches to chatbot creation. *Journal of Artificial Intelligence Research*, v.65, p.123-145, 2022.

- Brown, T.; et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, v.33, p.1877-1901, 2020.
- Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, p.4171-4186, 2019.
- Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, p.328-339, 2018.
- Klemmer, E.; et al. Evaluating Voice-based Assistant for Visually Impaired Users. *Proceedings of the ACM on Human-Computer Interaction*, v.4, n.CSCW2, p.1-23, 2020.
- LM Studio. Discover, download, and run local LLMs. Disponível em: <https://lmstudio.com>. Acesso em: 18 set. 2023.
- nickolaslivero/phidata. Phidata Repository. Disponível em: <https://github.com/nickolaslivero/phidata>. Acesso em: 18 set. 2023.
- Phidata. LLM OS Architecture. Disponível em: <https://github.com/phidatahq/phidata>. Acesso em: 18 set. 2023.
- Radford, A.; et al. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 2019.
- Rafat, M. I. AI-powered Legal Virtual Assistant: Utilizing LLM Optimized by RAG for Housing Dispute Resolution in Finland. *Artificial Intelligence and Law*, v.31, p.67-88, 2023.
- Simeoni, I.; Torroni, P. Empathic Voice: Enabling Emotional Intelligence in Virtual Assistants. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p.1021-1032, 2021.
- Touvron, H.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.
- Vaswani, A.; et al. Attention Is All You Need. In: *Advances in Neural Information Processing Systems*, v.30, p.5998-6008, 2017.
- Vu, M. D.; et al. GPTVoiceTasker: AI-Powered Voice Assistant for Smartphones. *Mobile Computing and Communications Review*, v.27, n.3, p.44-59, 2023.