Assistente de voz integrado com *Large Language Model*

1st João Felipe Gobeti Calenzani
Departamento de Informática UFES
Vitória, Brasil

2nd Victor Nascimento Neves
Departamento de Informática UFES
Vitória, Brasil

3rd Lauro José Lyrio Júnior
Departamento de Informática UFES
Vitória, Brasil

***Abstract*—Este artigo apresenta um sistema de chat interativo que integra texto e voz, utilizando APIs para conectar o reconhecimento de voz do modelo SeamlessM4T e a capacidade de inferência do modelo Llama 2. O sistema se adapta a múltiplos idiomas, com o modelo StyleTTS2 sendo utilizado para sintetizar respostas em inglês. O frontend, desenvolvido com React, gerencia o estado e a asincronicidade das interações, proporcionando uma interface de usuário coesa e responsiva. A arquitetura projetada permite uma comunicação fluída e natural, evidenciando o potencial dos assistentes virtuais suportados por IA no engajamento com o usuário em um contexto multilíngue.**

***Index Terms*—Chat Interativo; Reconhecimento de Voz; Processamento de Linguagem Natural; Síntese de Fala; Assistente Virtual;**

I. Introdução

Este artigo descreve a concepção e execução de um sistema interativo de chat que oferece aos usuários a capacidade de comunicar-se por meio de texto e voz. A plataforma é sustentada por uma arquitetura de backend robusta que integra dois modelos de inteligência artificial avançados através de requisições API pela plataforma replicate [1]: o LLM (*Large Language Model*) Llama 2 [2] para processamento de linguagem natural e o SeamlessM4T [3] para reconhecimento e síntese de voz. O modelo SeamlessM4T é responsável pela transcrição de áudio para texto, permitindo que a entrada vocal dos usuários seja convertida em um formato textual. Este texto é então fornecido ao modelo Llama 2, que processa a mensagem e produz uma resposta contextual relevante. Essa resposta é então processada pelo modelo SeamlessM4T para geração de um áudio que represente seu conteúdo textual, à exceção da língua inglesa, na qual foi utilizado o modelo StyleTTS2 [4]. O sistema fornece suporte a mais de 30 idiomas.

A infraestrutura do sistema foi cuidadosamente desenhada para suportar a troca dinâmica entre os modos de texto e voz, garantindo que a experiência do usuário permaneça contínua independentemente do método de entrada escolhido. A implementação efetiva do frontend, utilizando o framework React, aborda desafios técnicos como gerenciamento de estado e atualizações assíncronas, garantindo que as interações do usuário sejam refletidas na interface com responsividade e precisão.

Identify applicable funding agency here. If none, delete this.

Este estudo também discute as complexidades envolvidas nas requisições concorrentes a APIs externas e apresenta as estratégias adotadas para manter a estabilidade e a eficiência do chat. A pesquisa contribui significativamente para o campo de interfaces conversacionais, demonstrando como a combinação de modelos de IA distintos pode resultar em uma comunicação efetiva e natural entre humanos e assistentes virtuais.

II. Trabalho Proposto

Este trabalho consiste na integração de modelos de inteligência artificial através de APIs, criando um sistema de chat híbrido que oferece interatividade por texto e voz. A abordagem metodológica é dividida em várias etapas fundamentais que garantem uma implementação eficaz e eficiente do sistema proposto.

A. Desenvolvimento Frontend:

Utilizando o framework React, desenvolvemos uma interface de usuário reativa e dinâmica. A gestão do estado foi meticulosamente projetada para lidar com atualizações assíncronas, garantindo que a interface refletisse o estado mais recente das interações do usuário. Componentes funcionais React foram empregados para renderizar mensagens de texto e controles de áudio, permitindo aos usuários usufruirem de ambos os modos de entrada de forma intuitiva.

B. Integração de APIs:

A comunicação com o modelo SeamlessM4T foi estabelecida para transcrever a entrada de áudio do usuário em texto. Posteriormente, esse texto foi enviado ao modelo Llama 2 via API para gerar uma resposta contextualizada. A integração foi facilitada por meio de funções assíncronas que garantiam a execução de chamadas de API sem bloquear a interface do usuário.

C. Modelo Llama 2:

Foi incorporado o modelo de linguagem Llama 2 como um componente chave para interpretação e resposta textual. Com sua arquitetura de rede neural sofisticada, disponível em variantes de 7 bilhões, 13 bilhões ou 70 bilhões de parâmetros, o modelo foi escolhido por sua capacidade superior de compreensão e geração de linguagem natural. As requisições ao Llama foram estruturadas para operar em modo de streaming, permitindo que as respostas fossem geradas e transmitidas

de forma contínua e em tempo real. Isso proporcionou uma interação mais fluida e uma redução significativa na latência, essencial para a conversação interativa em um chat.

D. Modelo Seamless M4T:

Em contrapartida, o emprego do modelo Seamless M4T foi bifacetado, destacando-se notavelmente na tarefa de transcrição de áudio para texto. A precisão na conversão demonstrou a robustez do modelo em entender e transcrever com fidelidade a voz capturada, o que foi essencial para a inicialização do diálogo no sistema de chat. No entanto, a funcionalidade de síntese de texto para voz enfrentou desafios substanciais. O modelo impunha uma limitação na duração dos áudios gerados, restringindo-os a um máximo de 20 segundos. Esta restrição afetava adversamente tanto o conteúdo quanto a qualidade da saída de áudio, pois frases mais longas ou complexas não eram totalmente capturadas na síntese, fragmentando assim a experiência conversacional.

E. Modelo StyleTTS2:

A escolha do modelo StyleTTS2 para lidar com a síntese de voz em inglês surgiu como uma medida para superar as limitações encontradas no M4T. O StyleTTS2 foi capaz de gerar áudios mais longos e com uma qualidade de voz mais natural e contínua, o que se alinhava melhor com as expectativas de interações mais extensas e detalhadas em inglês. A transição para o StyleTTS2 refletiu a busca por aprimoramento na qualidade e na fluência da comunicação, priorizando a entrega de uma experiência de usuário mais rica e engajadora.

F. Captação e Processamento de Áudio:

A captação de áudio é realizada através do microfone do usuário, utilizando APIs de mídia do navegador. Após a captura, o áudio é processado e codificado em base64, um formato compatível para a transmissão via HTTP. Essa codificação é crucial para encapsular o conteúdo binário do áudio, permitindo que o dado seja enviado de forma segura e eficiente para a API do modelo Seamless M4T.

G. Personalização Linguística:

A interface do usuário foi desenhada para permitir a seleção da língua desejada, proporcionando uma experiência customizada e acessível. Dependendo da língua escolhida, o sistema ajustava dinamicamente o modelo de IA utilizado para a geração de áudio, com o modelo StyleTTS2 sendo utilizado especificamente para respostas em inglês, garantindo assim, uma resposta sintetizada mais natural e apropriada para o idioma selecionado.

H. Estratégias de Requisição:

Devido à natureza assíncrona das chamadas de API e à necessidade de um diálogo contínuo, adotamos estratégias de requisição que incluíam a verificação do estado de processamento e a obtenção de resultados de predição. Isso permitiu que o sistema respondesse de maneira eficiente, mesmo diante de operações de longa duração, como a geração de áudio.

III. Experimentos e Resultados

A. Configuração e Execução dos Testes:

Os testes foram configurados para explorar as funcionalidades do sistema, incluindo seleção do modelo de IA, definição de linguagem, ajuste do prompt inicial e calibração de parâmetros como temperatura, top_p e max_tokens (ver Figura 1). As interações foram avaliadas quanto à precisão na transcrição de voz para texto, à relevância das respostas geradas pelo modelo Llama e à qualidade da síntese de voz.

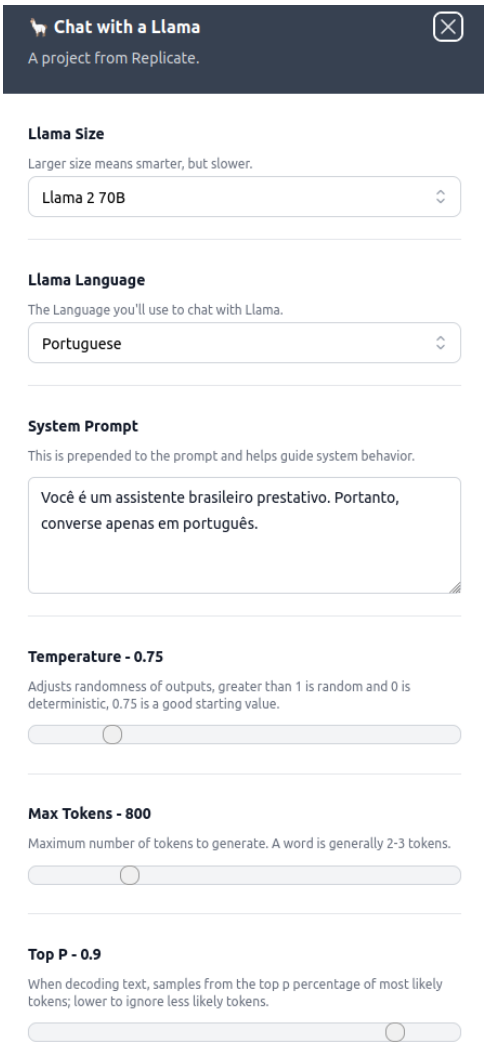

Fig. 1. Painel de configuração.

B. Resultados e Análise:

Os resultados obtidos demonstraram a capacidade do sistema de proporcionar uma comunicação eficaz, embora tenham sido identificadas áreas críticas para aprimoramento. A limitação na geração de áudio pelo modelo M4T, restrita a clipes de até 20 segundos, representou um desafio significativo, principalmente para idiomas que não o inglês. A implementação do modelo StyleTTS2 mitigou essa limitação para usuários de língua inglesa, mas evidenciou a necessidade

de soluções semelhantes para outros idiomas. Os testes podem ser visualizados nos vídeos disponibilizados no repositório do github. Uma captura de tela de uma execução pode ser vista na figura abaixo:

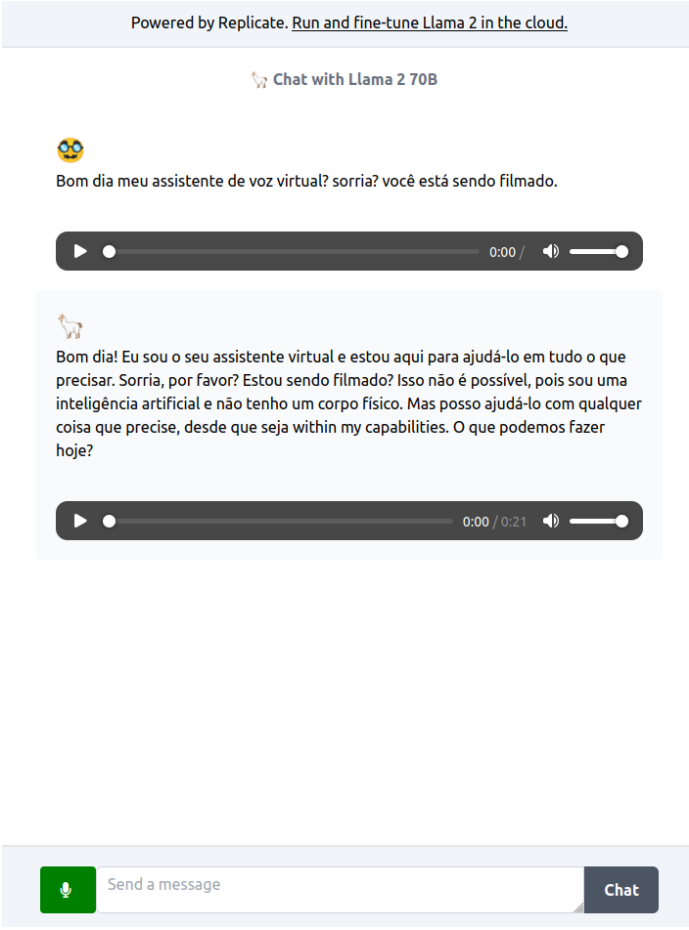

Fig. 2. Exemplo de uso.

C. Perspectivas de Melhoria:

A pesquisa sugere que um foco imediato para melhoria deveria ser a expansão das capacidades do modelo de geração de áudio, especialmente para idiomas além do inglês. A busca por modelos alternativos que possam oferecer síntese de voz de maior duração e com qualidade superior é crucial para aprimorar a experiência do usuário. Uma alternativa é o próprio modelo M4T em uma versão mais recente, que só não foi explorada nesse trabalho pela inviabilidade de, num primeiro momento, utilizar esse modelo via API.

Além disso, a implementação de funcionalidades de streaming para a síntese de voz emergiu como uma área promissora para desenvolvimento futuro. A viabilização de streaming na geração de áudio pode aumentar significativamente a fluidez da conversa, aproximando ainda mais a experiência de interação com um assistente virtual daquela de uma conversa com um ser humano. Isso não só melhoraria a continuidade do diálogo, mas também aumentaria a naturalidade e a imersão nas interações.

D. Conclusão:

Os experimentos realizados confirmam o potencial do sistema de chat interativo como uma ferramenta avançada de interação humano-computador. As áreas identificadas para melhorias fornecem uma direção clara para futuras pesquisas e desenvolvimento, com o objetivo de alcançar uma experiência de usuário mais rica e envolvente. A integração de tecnologias de IA de ponta e a adaptação contínua às necessidades dos usuários permanecem como elementos centrais para o avanço nesse campo.

References

[1] Replicate, "Replicate." [Online]. Available: https://replicate.com/

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[3] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P.-J. Chen, N. E. Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M.-J. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. Peloquin, M. Ramadan, A. Ramakrishnan, A. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. C. jussà [3], O. . andCelebi andMaha Elbayad andCynthia Gao, F. Guzmán, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, and S. Wang, "Seamlessm4t—massively multilingual multimodal machine translation," *ArXiv*, 2023.

[4] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *ArXiv*, vol. abs/2306.07691, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259145293