Evaluation Framework and Benchmark for Multi-lingual LLM Voice Assistants

Madhumitha Sivalingapandian*
Brentwood High School
Brentwood, TN 37027
mithu.sivali@gmail.com

Abstract

There has been a lot of development in Large Language models (LLMs), which has created a new era of voice assistants capable of understanding and interacting across multiple languages. However, evaluating the performance and effectiveness of these multilingual voice assistants remains a significant challenge. This paper presents a comprehensive benchmarking framework designed to assess the capabilities of cascaded voice chat models across different languages. This framework is grounded in a real-world study conducted in Pattiveeranpatti, a village in India. Helping school children learn technologies such as embedded software. This framework has been used as a benchmark for the practical applications of voice based assistants for this population.

The importance of this work lies in its potential to standardize the evaluation process for multilingual LLM based voice assistants, offering a clear set of metrics and methodologies that can be universally applied. By focusing on both technical performance and user experience, our framework aims to identify strengths and weaknesses in current models, guiding future improvements and innovations. Furthermore, the study highlights the critical role of voice assistants in enhancing accessibility and inclusivity, particularly in underrepresented and multilingual communities. This research not only contributes to the academic discourse on natural language processing but also has immense practical implications for developers and policymakers aiming to create more effective and equitable AI-driven communication tools.

1 Background

Pattiveeranpatti is a village near South India. Where, I've taught students between the grades of 6th - 8th at the Metric School and All Girls School.[9] [8] I've taught 15 students at the Metric School and 30 at the All Girls School. There are significant language issue when I was teaching at both schools. The language barrier was significant that even with people to translate, it was still quite difficult, despite me knowing Tamil as well as them knowing English. But it is not enough to form coherent sentences and understand me if I speak English. [5]

We've brought a total of 20 laptops and 45 Micro: Bits to the schools so that the children can learn better and be less digitally divided [12]. This was in partnership with my Girl Scout Gold Award Project where we collected a donation of 20 laptops from Medhost and 45 Micro: Bits from people who read about the cause of this project. Since Pattiveeranpatti is a very small village, the power cuts off a lot. We collected laptops so that work could be done even if the power went out. The Micro: Bits

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

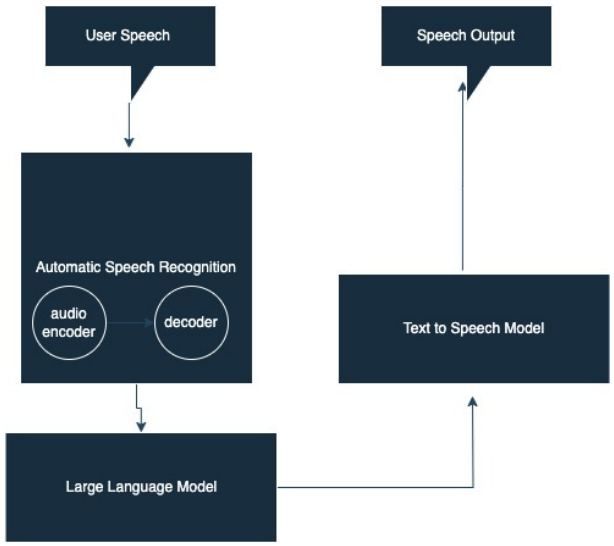

Figure 1: Cascaded voice chat model

was so that the students would learn Python more easily and be interested in programming. We've also provided each of them with a curriculum as well as Python projects that they could do on their own time. There was still the language barrier issue but we overcame it and it was not as much of an issue as it was last year.

One of the differences we noticed was when we were explaining the concept of Rock, Paper, and Scissors to them. When we first told them about the game they did not understand what we meant, but then they realized we meant Stone, Paper, Scissors. To them, it was Stone instead of Rock which was interesting because even though their name for the game was in English as well, it was different. They also could not relate Stone as a synonym to Rock. These small variations in language make it difficult for each of us to understand each other.

2 Literature Review

There are prior work evaluating language models, for example work by [14] evaluated multilingual BERT model [2]. In recent years there have been many large language models that were built which have hundreds of billions in parameters. Parameters are individual connections in a Machine learning model that enables the model to match patterns in text. With a larger parameter count, these language models are shown to have emergent properties such as semantic understanding of different languages.

3 Cascaded voice chat models

The process we went through to build our systems are called a cascaded system [3]. We first took a speech sample in waveform and put it through an Automatic Speech Recognition (ASR) Model also known as ASR. This step takes the audio and converts it into text form. Then, this text output is then sent to a Large Language Model also known as LLM. This step takes a model with large amounts of data which allows them to generate natural language to perform tasks like answering questions. It answers the question and then takes the response in text form and puts it through a Text-to-Speech model also known as TTS. This final step converts the text into waveform speech and then we evaluate it.

4 Proposed Evaluation Framework

Our evaluation framework for multilingual LLM voice assistants comprises three key components: Automatic Speech Recognition (ASR)[13], Language Understanding (LU), and Text-to-Speech

	literature	grammar	civics	econ	geo	chem	physics	math
llama-8B	20%	28.7%	25%	35%	27%	21%	20 %	19.4 %
gpt-4o	37.5%	21.2%	31.25	22.5%	37.5%	28.75%	23.75%	37.1%

Table 1: Language Understanding benchmark results

(TTS). Each component is measured against specific performance metrics to provide a comprehensive assessment of the voice assistant's capabilities.

4.1 Automatic Speech Recognition (ASR)

This is the speech recognition part where we evaluated the model's multi-lingual language understanding ability. Specifically in this case, we chose Tamil because this is the language that is used in the schools that we were teaching to. We created a corpus of Tamil speech text which we utilized to benchmark.

Word Error Rate (WER): This metric quantifies the accuracy of the ASR model by comparing the transcribed text to the reference text. [10] Lower WER indicates higher accuracy.

Latency: Measures the time taken from the end of the user's speech to the completion of the transcription. Lower latency indicates a more responsive ASR system. We've created a benchmark dataset of. This is important as we deal with rural areas, latencies add up and results in poor performance leading to lack of interest.

4.2 Language Understanding (LU)

Language Understanding enables the system to understand the words that are spoken and then convert them into a useful answers , such as an explanation, summarization or conversation. As the name implies These language models are huge and require a lot of resources to run. In our case we wanted it to be accessible in remote areas where network coverage could be very low. In those cases it would be great have a model that can run locally and still produce responses.

We created the largest language understanding dataset with over 1000 questions[6]. This benchmark was based on high school questions as this application targeted a school audience. similar to MMLU [4] but with multilinguality in mind. This dataset is the largest of its type for this model

With this dataset we evaluated llama-3 as well as gpt-4o models. Based on this result shown in table 1 we utilized gpt-4o as a stepping stone before we finetune the model

Language Understanding Accuracy: Evaluates the model's ability to correctly interpret and respond to user queries across different languages. This involves assessing the precision of intent recognition and entity extraction.

Size of the model We evaluate the size of the model as this lets us predict whether a model could be deployed offline or needs to be online perpetually. A smaller model can be run on a smaller machine.

Time to First Token (TTFT): This is where we measure the time taken for the LLM to generate the first token of the response after receiving the input. Lower TTFT indicates a faster and more efficient language model.

4.3 Text-to-Speech (TTS)

Evaluating speech is difficult since human speech has so many nuances such as emotion in delivering, diction, and much more, it becomes hard to figure out benchmarks to judge the accuracy of such a human-like quality. A typical metric used for audio quality is **Mean Opinion Score**[1]. This is where we took three people and had them judge the outputs of 20 responses to questions from 3 different Text-to-Speech Models. The Text-to-Speech Models are an open source model called llama7b, and other proprietary models from openai-tts, and elevenlabs.

Latency: Measures the time taken from the end of the language model's response generation to the completion of the speech synthesis. Lower latency indicates a faster TTS system. We found all three

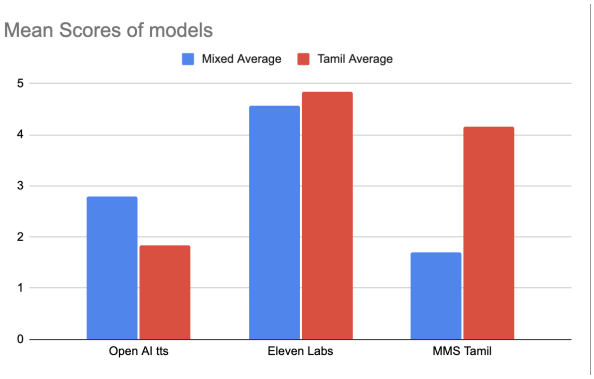

Figure 2: Text to Speech mean opinion score

systems where within <500 ms limit. We intend to test this complete on our next visit, so that we can understand how latency gets impacted on ground.

based on Figure 2 we decided utilize elevenlabs as they had the highest accuracy. One interesting point is that, the MMS model [7] performed better when the text had only Tamil words. When mixed with English, it was unable to perform well. This is a disadvantage of models that can only handle one language.

By systematically evaluating these components, our framework provides a robust method to benchmark and improve the performance of multilingual LLM voice assistants, ensuring they are both accurate and efficient in diverse linguistic settings.

5 Tools utilized

This was a great project that taught us to utilize a lot of tools that were interesting to work with. For all the analysis we utilized python with its pandas, torch, and other Machine Learning libraries, these can be found in the repository [11]. For running the machine learning models locally we used repositories from research instituitions such as facebook-research and others. We've kept all the code, with its changes at this github repo [11]. The dataset for the llm can be found here [6].

6 conclusion

Using our evaluation benchmarks, we were able to design a better voice assistant based on the data that we collected. This study highlights the importance of evaluation scores and benchmarks when dealing with machine learning models as they can sometimes be unpredictable.

Our Evaluation Framework paves the way for better measurement and design of multilingual voice assistants. We believe that with these machine learning model can reduce the digital divide and would empower and educate people from even remote parts of the world.

With our contribution to the language understanding benchmark for Tamil. We believe newer and better systems can be built and evaluated with ease.

Acknowledgments and Disclosure of Funding

I would like to extend my sincerest gratitude towards Suriya Ganesh Ayyamperumal who made this paper possible. He brought to my attention this opportunity and helped me collect more information on this subject. He mentored me and helped me debug. I would also like to give a special thanks to Ram Prakash who helped to annotate the outputs. None of this would have been possible without him.

References

[1] Min Chu and Hu Peng. An objective measure for estimating mos of synthesized speech. In *INTERSPEECH*, pages 2087–2090, 2001.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5792–5793, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599572. URL https://doi.org/10.1145/3580305.3599572.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[5] MadhuMitha Sivalingapandian. Video on youtube, 2024. URL GirlScoutGoldAwardProjectPVPCoders(DIGITALDIVIDE). Accessed: 2024-06-27.

[6] MithuSi. Multi-lingual llm dataset. https://huggingface.co/datasets/MithuSi/multi-lingual-llm, 2024. Accessed: 2024-06-27.

[7] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv*, 2023.

[8] PVPCoders. July 25, 2023 - nadar sundara visalatchi vidyasala matriculation higher secondary school, 2023. URL https://pvpcoders.org/2024/02/28/july-25-2023-nadar-sundara-visalatchi-vidyasala-matriculation-higher-secondary-school/. Accessed: 2024-06-27.

[9] PVPCoders. March 14, 2024 - nadar sundara visalatchi vidyasala matriculation higher secondary school, 2024. URL https://pvpcoders.org/2024/06/27/march-14-2024-nadar-sundara-visalatchi-vidyasala-matriculation-higher-secondary-school/. Accessed: 2024-06-27.

[10] Matt Shannon. Optimizing expected word error rate via sampling for speech recognition. *arXiv preprint arXiv:1706.02776*, 2017.

[11] Mithu Sivali. Multi-lingual llm. https://github.com/mithu-sivali/multi-lingual-llm, 2024. Accessed: 2024-06-27.

[12] Madhumitha Sivalingapandian. Girl scout gold award project execution pvp coders (digital divide), 2024. URL https://www.youtube.com/watch?v=k4xgL6RL1V4. Accessed: 2024-04-27.

[13] Dong Wang, Xiaodong Wang, and Shaohe Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018, 2019.

[14] Daniel Yue Zhang, Jonathan Hueser, Yao Li, and Sarah Campbell. Language-agnostic and language-aware multilingual natural language understanding for large-scale intelligent voice assistant application. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1523–1532, 2021. doi: 10.1109/BigData52589.2021.9671571.