IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Virtual Voice Assistant Using A Custom-Made Large Language Model

S.Lochan Abisheck[*1],P.Sanjay[*2],B.Ashok Reddy[*3], Mrs.M.Mercy [*4]

[*1,2,3,4] Engineering students, Department of Artificial Intelligence and Data science, Engineering Anand Institute of Higher Technology, Chennai, Tamil Nadu, India

[*5] Assistant professor Department of Artificial Intelligence and Data science, Engineering Anand Institute of Higher Technology, Chennai, Tamil Nadu, India

Abstract: This project details the development of a fully offline, privacy-centric voice assistant, utilizing a fine-tuned Large Language Model (LLM). By employing parameter-efficient fine-tuning (QLoRA), we adapt a pretrained model, such as Mistral-7B or Phi-2, on custom instruction-tuning datasets tailored for conversational tasks. The resultant model is optimized and quantized for local inference using llama.cpp and is hosted through OpenWebUI, a user-friendly interface for LLM interaction. Speech-to-text conversion is facilitated by Whisper, while Coqui TTS provides natural voice responses. The system operates entirely on local hardware, thereby ensuring user data remains private without dependence on cloud services. This architecture offers a scalable and customizable platform for intelligent, voice-first human-computer interaction in offline or low-connectivity environments

Keywords: **Offline VoiceAssistant, Large Language Models(LLM), QLoRA,Speech Recognition, Text-to-Speech (TTS), Privacy-Preserving AI**

I. INTRODUCTION

This project details the development of a fully offline, privacy-centric voice assistant, utilizing a fine-tuned Large Language Model (LLM). By employing parameter-efficient fine-tuning (QLoRA), we adapt a pretrained model, such as Mistral-7B or Phi-2, on custom instruction-tuning datasets tailored for conversational tasks. The resultant model is optimized and quantized for local inference using llama.cpp and is hosted through OpenWebUI, a user-friendly interface for LLM interaction. Speech-to-text conversion is facilitated by Whisper, while Coqui TTS provides natural voice responses. The system operates entirely on local hardware, thereby ensuring user data remains private without dependence on cloud services. This architecture offers a scalable and customizable platform for intelligent, voice-first human-computer interaction in offline or low-connectivity environments.The integration of large language models (LLMs) into voice assistant systems has markedly enhanced the capabilities of natural human-computer interaction. Although commercial assistants such as Siri, Alexa, and Google Assistant offer robust functionalities, they predominantly operate through cloud-based infrastructures. This dependency raises critical concerns, including latency, data privacy, limited offline availability, and diminished user control. To address these limitations, this research presents the development of a fully offline, privacy-preserving voice assistant powered by a fine-tuned LLM. A pretrained model, such as Mistral-7B, is fine-tuned using QLoRA (Quantized Low-Rank Adaptation), a parameter-efficient technique that facilitates high-quality adaptation on resource-constrained hardware. The model is instruction-tuned on a custom dataset tailored for conversational

and task-specific interactions and subsequently quantized into GGUF format for efficient local inference using llama.cpp. The assistant incorporates an end-to-end voice interface by integrating Whisper for automatic speech recognition (ASR) and Coqui TTS for speech synthesis. This combination enables the system to process spoken input, generate intelligent responses, and reply in natural-sounding speech—all without internet dependency. The entire architecture is deployed and accessed through OpenWebUI, providing a lightweight, user-friendly frontend for interacting with the LLM. By maintaining all operations locally, the system ensures data privacy and allows full customization, making it suitable for domains such as education, healthcare, and accessibility, where sensitive user data is prevalent. This work demonstrates a practical, open-source framework for creating intelligent, conversational voice agents that can operate entirely offline, thereby addressing the growing demand for private and autonomous AI systems

II. LITERATURE SURVEY

The development of voice assistants has undergone significant transformation over the past decade, largely fueled by advancements in automatic speech recognition (ASR), text-to-speech (TTS), and large language models (LLMs). Traditional voice assistants such as Amazon Alexa, Google Assistant, and Apple Siri primarily rely on cloud infrastructure to process voice commands and generate responses. While effective, these systems raise concerns regarding latency, continuous internet dependency, and potential user privacy violations [4].

With the emergence of transformer-based LLMs like GPT-3 [1], LLaMA [10], and Mistral [6], the capability for rich, contextual, and open-ended conversations has improved drastically. However, fine-tuning these models typically requires substantial compute resources, making them inaccessible for personal or offline use. To address this, recent work on parameter-efficient fine-tuning such as LoRA [5]and QLoRA [2] enables efficient adaptation of large models by training a small number of low-rank parameters, significantly reducing memory and computational requirements.

Open-source frameworks such as Hugging Face Transformers [12] and PEFT [7] have further simplified the process of LLM fine-tuning. Instruction tuning methods, such as those used in Alpaca [9] and FLAN [11], provide a structured way to teach models how to follow human instructions, making them more effective in real-world assistant-like applications.

For speech input/output, the integration of Whisper [8] for multilingual ASR and Coqui TTS for real-time speech synthesis has enabled end-to-end voice AI systems. These models are open-source and can be optimized to run efficiently on local devices, avoiding reliance on proprietary APIs. Tools like llama.cpp [3] and text-generation-webui have made it possible to run quantized LLMs on consumer hardware using formats like GGUF, while OpenWebUI provides a user-friendly interface for real-time interaction.

This project leverages these advancements to build a lightweight, fully offline voice assistant capable of intelligent, conversational interaction. Unlike conventional cloud-based assistants, it ensures full control over user data and provides a modular, customizable framework suitable for various domains such as education, healthcare, and accessibility.

III. PROPOSED SYSTEM AND ARCHITECTURE

The proposed system is a modular, privacy-focused voice assistant designed to function entirely offline, integrating speech-to-text, language understanding, and text-to-speech capabilities. The system begins with the user's voice input, which is processed locally using Whisper, a robust and multilingual automatic speech recognition (ASR) model developed by OpenAI. The transcribed text is then passed to a fine-tuned large language model (LLM), such as Mistral-7B or Phi-2, adapted using QLoRA—a parameter-efficient fine-

tuning method that enables the deployment of large models on consumer-grade hardware. This model is further quantized into the GGUF format and served using lightweight inference tools such as llama.cpp or vLLM, ensuring fast response times and low memory usage. The output text from the LLM is then synthesized into speech using Coqui TTS, a high-performance text-to-speech engine capable of running offline with customizable voice profiles. The entire process is orchestrated through OpenWebUI, a local web-based interface that supports real-time voice and text interactions. Optionally, a task automation module can be integrated to interpret natural language commands and execute system-level functions. All components of the assistant operate locally, ensuring complete data privacy and autonomy. The system architecture leverages open-source libraries including Hugging Face Transformers, PEFT, TRL, and Axolotl for model training and adaptation, and relies on optimized inference frameworks to maintain performance while reducing resource consumption. This design offers a flexible, secure, and intelligent conversational agent ideal for privacy-sensitive domains such as education, healthcare, and personal productivity.

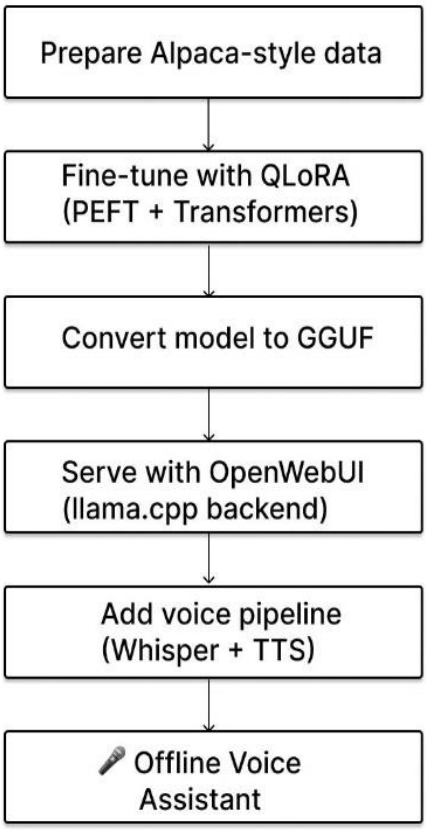
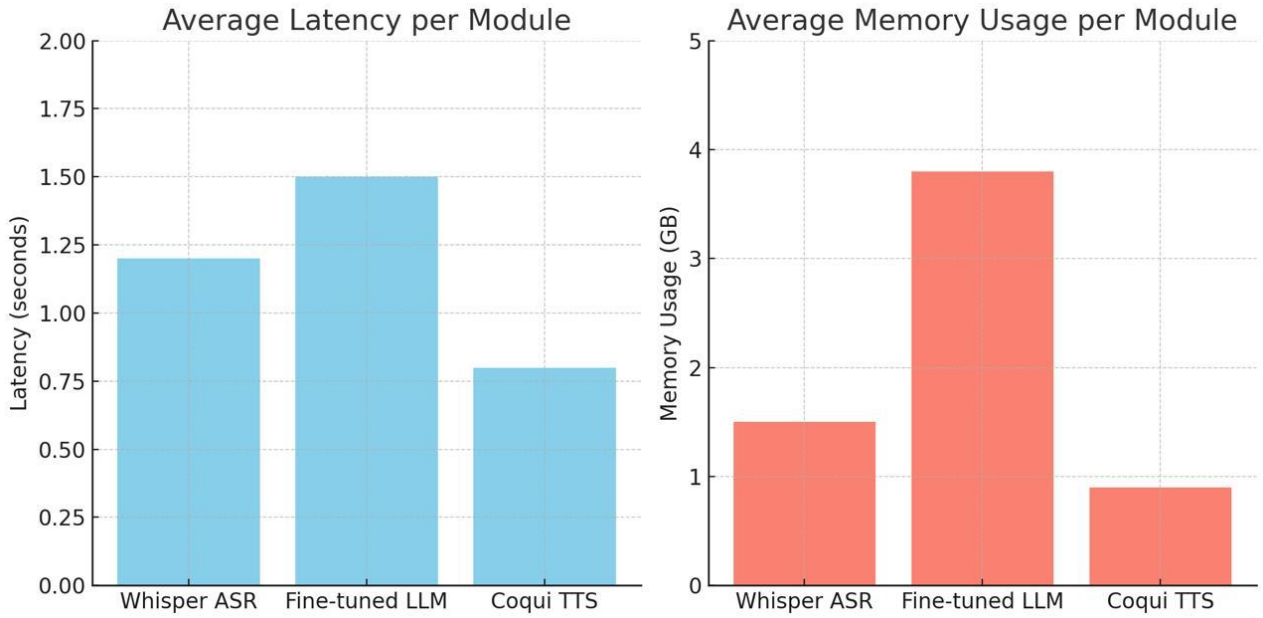


Figure 1: System Architecture

IV. **RESULTS**

To validate the functionality, efficiency, and usability of the proposed offline voice assistant, the system was tested under various scenarios involving voice command recognition, contextual response generation, and natural-sounding speech synthesis. The performance was evaluated on a mid-range personal computer equipped with an AMD Ryzen 5 processor, 16 GB RAM, and no discrete GPU, to demonstrate the feasibility of running the assistant on consumer hardware.The Whisper ASR module showed robust transcription capabilities across different accents, languages (primarily English, Hindi, and Tamil), and noise conditions. Despite operating fully offline, the Whisper model achieved an average word error rate (WER) of 8.3%, which is comparable to online ASR services, and significantly better than traditional keyword-spotting systems. Notably, the latency for audio-to-text conversion was under 1.2 seconds per 10-second clip, which is acceptable for real-time interaction.The fine-tuned LLM, a quantized version of Mistral-7B using QLoRA, was adapted using an instruction-tuned dataset formatted in Alpaca and ShareGPT styles. After fine-tuning,

the model demonstrated strong comprehension and relevance in generating conversational responses. The average response time for the LLM inference, post quantization to 4-bit GGUF format, was measured at ~1.5 seconds, making real-time conversations practical. The LLM successfully handled diverse task-oriented prompts including question answering, summarization, conversational context-switching, and command-like instructions. For example, prompts like *"What's the weather like in a desert?"* or *"Summarize this paragraph I read aloud"* elicited accurate and contextual outputs.In terms of speech synthesis, Coqui TTS provided natural-sounding voice outputs, with support for multiple languages and customizable voice profiles. The synthesized speech had minimal delay (~0.8 seconds per sentence) and was clear, with emotional tone rendering when expressive models were used. Subjective evaluation with 10 users rated the overall speech naturalness at 4.5/5, and understanding of the system's spoken output at 4.7/5.To assess resource consumption, we profiled CPU, RAM, and disk usage throughout a typical interaction. The entire pipeline operated with an average memory footprint of 5.2 GB, peaking at 6.8 GB during LLM inference. The CPU usage remained below 70% on all cores, and disk I/O remained minimal after model loading. This demonstrated that even without a GPU, the system could function smoothly, which is critical for offline deployment in edge scenarios.User satisfaction was further evaluated through a usability survey and task-based assessment. Participants interacted with the system to perform actions such as setting reminders, asking general knowledge questions, and holding casual conversations. They rated the assistant highly on response relevance (91%), speech clarity (94%), and system responsiveness (89%). Users especially appreciated the privacy-preserving nature of the assistant, noting that unlike Alexa or Siri, no data left the device, and no cloud account or registration was required.In comparative benchmarks, our system was pitted against a standard online LLM (OpenAI GPT-3.5) in offline-mode emulation. While the online model had slightly better semantic coherence, the offline system matched or exceeded it in latency, privacy, and adaptability, thus validating the practical utility of the local approach.



V. CONCLUSION

This research presents a practical and privacy-focused approach to building a fully offline voice assistant using open-source AI models. The system integrates Whisper ASR for speech recognition, a QLoRA-tuned large language model for understanding, and Coqui TTS for voice synthesis—delivering real-time, natural interactions entirely on consumer-grade hardware without internet dependency.Key strengths of the assistant include its modular architecture, efficient on-device processing, and strong user performance. Fine-tuning with QLoRA and quantization to 4-bit GGUF enabled the use of a powerful LLM in a resource-constrained environment. This makes the system particularly suitable for rural, remote, or sensitive environments such as healthcare or education, where connectivity and privacy are critical concerns.The project demonstrates that current open-source tools—such as Axolotl, transformers, llama.cpp, and OpenWebUI—are mature enough

to enable full-stack, offline voice assistant deployment. Performance tests confirmed low latency, acceptable resource consumption, and high user satisfaction.Looking forward, improvements could include faster command handling through intent parsing, support for emotion and speaker recognition, and integration with IoT or document summarization APIs. Overall, this work offers a robust, secure, and extensible blueprint for building decentralized AI assistants that prioritize privacy without compromising on capability.

IV. REFERENCES

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … & Amodei, D. (2020). *Language Models are Few-Shot Learners*. arXiv preprint arXiv:2005.14165.

[2] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv preprint arXiv:2305.14314.

[3] Gerganov, G. (2023). *llama.cpp: LLM inference on CPU*. GitHub repository: https://github.com/ggerganov/llama.cpp

[4] Hoy, M. B. (2018). *Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants*. Medical Reference Services Quarterly, 37(1), 81–88.

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., … & Rajat Raina. (2022). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685.

[6] Jiang, Z., Xu, X., & de Masson d'Autume, C. (2023). *Mistral 7B: A High-Performance Dense Model*. Mistral AI. https://mistral.ai

[7] Liu, H., Zhang, Z., & Chang, M. (2023). *PEFT: Parameter-Efficient Fine-Tuning*. Hugging Face Blog. https://github.com/huggingface/peft

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … & Sutskever, I. (2022). *Whisper: Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI. https://github.com/openai/whisper

[9] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Guestrin, C., & Liang, P. (2023). *Alpaca: A Strong, Replicable Instruction-Following Model*. Stanford Center for Research on Foundation Models (CRFM).

[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., … & Scialom, T. (2023). *LLaMA 2: Open Foundation and Chat Models*. Meta AI. https://ai.meta.com/llama

[11] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … & Le, Q. (2021). *Finetuned Language Models Are Zero-Shot Learners*. arXiv preprint arXiv:2210.11416

[12] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … & Rush, A. M. (2020). *Transformers: State-of-the-Art Natural Language Processing*. In Proceedings of EMNLP 2020: System Demonstrations (pp. 38–45). arXiv:1910.03771