ConverseAI: Revolutionizing Edge AI and Vision with customizable models

Mohit Manjalkar
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
manjalkarmohit-cpmn@atharva.coe.ac.in

Sonal Baury
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
baurysonal-cmpn@atharvacoe.ac.in

Dhanush Chandran
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
chandrandhanush-cmpn@atharvacoe.ac.in

Sahil Bhojekar
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
bhojekarsahil-cpmn@atharva.coe.ac.in

Bhavna Arora
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
arorabhavna-cpmn@atharva.coe.ac.in

***Abstract*— ConverseAI is a robust, locally deployed artificial intelligence platform specifically designed to provide effortless interaction with optimized Large Language Models (LLMs), and its base engine is Ollama. With a focus on user privacy, personalization, and operation efficiency, it enables users to perform different tasks without the need for continuous internet connectivity. The web-based interface enables multimodal interaction in the form of text typing, vision-based inputs to handle images, and speech-to-text functionality for hands-free interaction. With Retrieval-Augmented Generation (RAG) built into it, ConverseAI enhances response precision with the use of document retrieval and embeddings to make responses relevant to the context. Its scalable design enables the deployment of various sizes of models and domain-specific variations, hence supporting a wide variety of workflows, which vary from educational support and professional automation to creating creative content. In successfully bridging the gap between artificial intelligence capabilities and real-world utilization, ConverseAI offers a very versatile and efficient AI experience.**

***Keywords*— large language models, customization, efficiency, offline ai, web-based interface, multimodal interactions, text prompts, image analysis, speech-to-text, hands-free usability, retrieval-augmented generation, document retrieval**

I. INTRODUCTION

In the past few years, the fast growth of Large Language Models (LLMs) has had a major impact on different industries, improving automation, access to knowledge, and intelligent assistance across various domains. Most AI-based platforms, however, are highly dependent on cloud services, and this has been a concern regarding privacy, internet dependency, and limited flexibility for customization. ConverseAI addresses these issues by providing a locally hostable LLM solution that facilitates users with seamless interaction ability while ensuring data security, customization, and efficiency. ConverseAI uses ollama as its backend engine to effectively handle fine LLMs on local devices and hence avoid the need for cloud connectivity. This architecture provides users with full control over their interactions and hence preserves privacy by minimizing dependency on external servers. Our solution provides a web-based interface that provides ability for multimodal interactions, allowing users to communicate with our LLMs through text inputs, vision-based processing for image analysis, and speech-to-text functionality for hands-free interactions. These functionalities of ConverseAI make it flexible and usable across numerous applications, ranging from personal aid, business automation, to educational support. Among the highlight features of ConverseAI is the presence of Retrieval-Augmented Generation (RAG), which boosts the accuracy of the output through the application of document retrieval and embedding-based search techniques. This ability allows the system to generate responses that are more contextually accurate by accessing external data sources and user-specified knowledge bases. In addition, ConverseAI is scalable, allowing users to deploy different model sizes according to their computational resources and to incorporate domain-specific customizations to make their LLMs

behavior more aligned to their requirements. With versatility and flexibility in mind, ConverseAI is a high-performance tool for educational applications, professional automation, creative content generation, and knowledge management. By bridging the gap between cutting-edge AI technology and real-world user requirements, the platform is designed to provide a high-performance, efficient, and personal AI experience that enables users to leverage the power of LLMs in a way that is most productive and effective to their workflows. ConverseAI provides a very user-friendly and intuitive, with an easy-to-use interface that proivides users with varying levels of technical expertise to interact with the system and leverage its capablities. The web-based interface simplifies interaction with AI, and users can input text, upload images for visual processing, or dictate prompts through speech-to-text processing. Unlike most cloud-based AI platforms that impose interaction limits or high subscriptions, ConverseAI offers full local control, with users having the ability to run AI models without restrictions while maintaining complete ownership of their private data. This makes it particularly valuable for organizations and individuals who prioritize data confidentiality, compliance, and operational control. Moreover, ConverseAI's modular architecture allows for seamless integration with many additional tools and third-party services through its API, making it more suitable for domain-specific workflows. Developers and companies can leverage our platform's extensibility to incorporate custom LLM fine-tuning, API integrations, and domain-specific datasets. It can be effectively used for academic research, cybersecurity analysis, content generation, or enterprise knowledge management, our solution offers the versatility to optimize AI-driven workflows based on specific needs. Furthermore, with support for scalable deployment, users can execute models of varying sizes depending on their hardware capabilities, ensuring optimal performance without imposing computational overhead.ConverseAI offers a privacy-focused, locally hosted LLM solution with multimodal processing, RAG integration, and scalability for various applications. By eleminating dependency on cloud resources, it ensures data security, customization, and functionality control. Its modular architecture allows for easy expansion and integration into various workflows.

II. REVIEW OF LITERATURE

The evolution of artificial intelligence (AI) and Large Language Models (LLMs) has led to significant advancements in human-computer interaction. Various AI platforms have been developed to enhance accessibility, performance, and user experience, with a focus on privacy, efficiency, and multimodal capabilities.

A. *Locally Hosted AI and Privacy Concerns*

Recent research emphasizes the growing demand for locally deployed artificial intelligence systems to help counter privacy concerns and minimize dependence on cloud infrastructures (Zhao et al., 2022). GPT-4, LLaMA, and Mistral are some of the tools that have proved that LLMs can be locally deployed without sacrificing efficiency (Brown et al., 2020). ConverseAI is moving in this direction by using Ollama as its backend engine, thus enabling users to process data locally without transmitting sensitive information to distant servers..

B. *Customization and Efficiency in LLMs*

Specialized use case tuning for LLMs has become a primary research interest. Task-specific tuning has been found to enhance model accuracy and contextual relevance (Gao et al., 2023). Retrieval-Augmented Generation (RAG) also enhances AI output by marrying document retrieval and embeddings to improve factual accuracy and reduce hallucinations (Lewis et al., 2021). ConverseAI employs RAG to provide users with accurate and contextually relevant answers and is a valuable resource for educational, professional, and creative applications

C. *Multimodal AI Interaction*

Multimodal AI interfaces' integration has been the subject of significant research, emphasizing their capability to enhance usability and user experience. Zhang et al. (2023) showed that with the integration of text, voice, and vision inputs, usability in AI varies extensively across broad applications. Multimodal input and interaction in ConverseAI is facilitated with the capability for hands-free use and increased usability through speech-to-text and image processing.

D. *Scalability and Deployment of AI Models*

SScalability is also a critical consideration in AI deployment since research has been investigating how to maximize model performance across different

computational platforms (Pope et al., 2022). Deploying different model sizes enables proper resource allocation and enhanced task performance. Scalable deployment is facilitated by ConverseAI's architecture, enabling it to be compatible with different hardware capabilities and user requirements.

E. AI for Real-World Applications

Academic research identifies the increasing importance of artificial intelligence in education, automation, and creative sectors. Learning platforms based on AI have been shown to enhance learning outcomes (Liu et al., 2023). Automation through the use of large language models has also been shown to enhance productivity in workplaces (Chen et al., 2021). ConverseAI leverages such advancements to enable multiple workflows, thus connecting AI capabilities to real-world applications.

F. Conclusion

The literature highlights the need for locally deployed, customizable, and multimodal AI platforms. ConverseAI meets these criteria as a privacy-centric, efficient, and scalable solution for AI. Through the integration of fine-tuned LLMs and state-of-the-art retrieval mechanisms with multimodal capabilities, ConverseAI is an addition to the continued development of AI applications in many areas.

III. PROPOSED SYSTEM

The ConverseAI architecture based on large language models (LLMs) is created for simple, localized, and privacy-conscious interaction, with a web interface along with sophisticated multimodal support. The frontend user interface is constructed using contemporary web technologies such as Javascript, Svelte, CSS, and Typescript. The users are allowed to input data in text form, voice (with the use of speech-to-text technology), and images. The system accepts textual inputs in real-time and allows speech-to-text conversion by employing APIs such as Whisper or other similar tools. Additionally, for audio output, text-to-speech capability is incorporated to offer vocal responses.

The backend module, written in Python and Shell, handles all the requests and connects the frontend interface to the model-serving infrastructure below. Central to the backend is Ollama, utilized for handling fine-tuned language models. Ollama is the processing unit that takes user input and gives output, seamlessly handling text-based as well as multimodal requests. The backend also includes Retrieval-Augmented Generation (RAG) to enhance the response accuracy. The process involves pulling relevant context from internal knowledge bases or vector databases like Pinecone or Weaviate, thus enriching the responses with real-time, contextually relevant data.

The speech processing integration is at the core of the system, with a speech-to-text module that translates spoken input into text, and a text-to-speech module that voices the output responses. This integrated system ensures local processing of all operations, hence reducing dependence on cloud services and enhancing user privacy. The modularity provides futureproofing for augmentation, including the integration of further multimodal inputs or further model fine-tuning capabilities.

A. Key Features

- The system is designed for local use, where users and organizations can manage data without remote servers. The setup offers greater privacy, improved data security, and compliance with regulations by not using cloud processing. Furthermore, offline processing of large language models provides fast inference while maintaining the control of sensitive data.

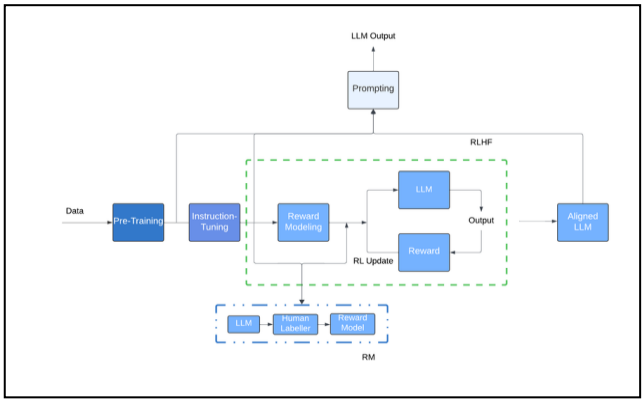

Fig. 1. Architecture of LLM Training, Fine-Tuning, and Alignment

- The platform supports multiple input modalities, including text, speech, and visual. The system can be used with natural language input, voice control, or by uploading images for analysis. The multimodal setup improves accessibility, user experience, and the application of artificial intelligence in different scenarios, including

voice assistants, document analysis, and image-based analysis.

- The system incorporates RAG to ensure generated output is more accurate and contextual. Through the retrieval of contextual information, knowledge bases, or relevant documents, the AI can generate well-informed, contextually sensitive responses. It is especially useful for business applications, research, and tailored support, keeping AI outputs accurate and contextual.
- Users can draw upon fine-tuned domain-specific large language models (LLMs) for specific industries, functionalities, or organizational use cases. The platform also provides model personalization, and users can enable improvements in areas such as health, finance, cybersecurity, education, and beyond. This provides organizations with the ability to make AI perform better by utilizing models that are industry-specific language- and procedure-aware.

B. *Hardware Requirements:*

- Processor: Multi-core 64-bit cpu with avx support.
- GPU : NVIDIA gpu with at least 8GB VRAM for efficient model inference.
- CUDA **:** If you plan to use GPUs for acceleration, you'll need to ensure CUDA is enabled and configured with the Nvidia Container Toolkit.
- Docker: Ensure you have Docker installed and running
- RAM: Minimum 16GB, recommended 32GB for larger models.
- Storage: 500GB SSD minimum, recommended 1TB SSD for faster performance.
- Audio/Camera: Microphone for speech-to-text and camera for vision tasks.

C. *Software Requirements:*

- Operating System: Linux (Ubuntu 22.04+), Windows 10/11 (64-bit), or macOS.
- Backend: Python 3.9+ with Ollama for model management.
- Frontend: Javascript, CSS, Svelte, Typescript.
- Speech-to-Text: Whisper for voice input.
- Computer Vision: OpenCV for image processing.
- Database: SQLite or PostgreSQL for data storage.
- Web Server: Nginx Proxy Manager for serving the platform.

IV. METHDOLOGY

The proposed system integrates a fine-tuned LLaMA model with a custom-built WebUI, enabling seamless interaction through text and speech-based inputs. The system operates within a Dockerized environment on a Linux-based machine, ensuring modularity and ease of deployment.

A. *System Architecture:*

The architecture consists of three core components:

1) Large Language Model (LLM) Processing – A fine-tuned LLaMA 3.2 8B model runs within an Ollama instance, handling all AI-driven responses.

2) Web Interface – A custom-built WebUI interacts with the user and forwards queries to the Ollama API, ensuring a responsive and intuitive experience.

3) Speech-to-Text Module – The system supports voice input, utilizing OpenDAI-Speech for high-accuracy speech-to-text conversion.

B. *Data Flow and Processing Pipeline:*

1) User Input Handling:

a) The user submits a query via the WebUI, either as text or voice input.

b) If voice input is provided, OpenDAI-Speech transcribes it into text.

2) Query Processing & Model Interaction:

a) The WebUI forwards the text input to the ollama instance via an API request.

b) The fine-tuned ConverseAI model processes the query and generates a response.

3) Response Generation & Delivery:

a) The AI-generated response is transmitted back to the WebUI.

b) The WebUI formats and displays the response for the user.

This pipeline ensures low-latency interactions while maintaining a modular and scalable design.

C. Algorithms and Models Used:

1) Large Language Model: Fine-tuned ConverseAI, optimized for domain-specific tasks.
2) Speech-to-Text: OpenDAI-Speech, utilizing Whisper-based transcription for high accuracy.
3) API-Based Communication: RESTful API interactions facilitate seamless data exchange between system components.

D. Implementation Details:

1) Backend: Ollama API for LLM processing.
2) Frontend: Custom WebUI with a modern, lightweight design for enhanced user experience.
3) Deployment: Docker containers for efficient resource management and system isolation.
4) Operating System: Fedora Linux provides a stable and secure environment for hosting the system.

V. RESULT ANALYSIS

The performance of **ConverseAI** was rigorously evaluated using a diverse set of standardized benchmarks designed to assess a broad spectrum of capabilities. These included general knowledge, coding proficiency, mathematics, logical reasoning, tool use, long-context understanding, and multilingual competence. In general knowledge, the model was tested on its ability to recall and synthesize factual information across various domains. Coding benchmarks measured its skills in generating, understanding, and debugging code across multiple programming languages. Mathematical tasks evaluated its ability to solve problems ranging from basic arithmetic to advanced algebra and calculus. The model's reasoning abilities were assessed through complex, multi-step problem-solving scenarios that required logical and analytical thinking. In tool use, ConverseAI was tested on its capacity to interact with external systems and APIs to extend its native capabilities. Long-context understanding tasks challenged the model to retain and reason over extended inputs, vital for applications involving large documents or sustained conversations. Multilingual benchmarks evaluated its fluency, comprehension, and translation abilities across several languages. The results demonstrate that ConverseAI delivers strong performance in reasoning, coding, and multilingual tasks, often outperforming or matching other state-of-the-art models. At the same time, the evaluations revealed areas for improvement, particularly in handling highly specialized queries and maintaining consistency in long-context reasoning. These insights are being used to guide future development and optimization of the model.

1) Performace Comparison Table

Benchmark	ConverseAI	Gemma 2 9B IT	Mistral 7B Instruct
MMLU	73.0	72.3	60.5
MMLU Pro	48.3	-	36.9
IFEval	80.4	73.6	57.6
HumanEval	72.6	54.3	40.2
MBPP EvalPro	72.8	71.7	49.5
GSMBK	84.5	76.7	53.2
MATH	51.9	44.3	13.0
ARC Challenge	83.4	87.6	74.2
GPQA	32.8	-	28.8
BFCL	76.1	-	60.4
Nexus	38.5	30.0	24.7
ZeroSCROLLS/QuALITY	81.0	-	-
InfiniteBench/En.MC	65.1	-	-
NH/Multi-needle	98.8	-	-
Multilingual MGSM	68.9	53.2	29.9

Performance Analysis

a) General Knowledge & Understanding:

i. ConverseAI achieves 73.0 on MMLU (0-shot, CoT), outperforming Mistral 7B Instruct (60.5) but slightly trailing Gemma 2 9B IT (72.3, 5-shot, non-CoT).

ii. In MMLU PRO, the model scores 48.3, which is lower than larger models like LLaMA 3.1 70B (66.4).

iii. The IFEval score of 80.4 indicates strong factual accuracy.

b) Coding Capabilities:

i. ConverseAI performs well in coding tasks, scoring 72.6 in HumanEval (0-shot) and 72.8 in MBPP EvalPlus (0-shot, base).

Fig. 2. User Interface

ii. It surpasses Mistral 7B Instruct (40.2, 49.5)

c) Mathematical Reasoning:

i. Strong performance in GSM8K (8-shot, CoT) with a score of 84.5, significantly higher than Mistral 7B Instruct (53.2) but lower than LLaMA 3.1 70B (95.1).

ii. MATH (0-shot, CoT) score of 51.9 suggests moderate mathematical capabilities, though it could be improved.

d) Logical & Reasoning Abilities:

i. ConverseAI achieves 83.4 in ARC Challenge (0-shot), showing strong logical reasoning skills.

ii. However, its GPOQA (0-shot, CoT) score is relatively low at 32.8, indicating challenges in complex reasoning.

e) Tool Use & Long-Context Handling:

i. ConverseAI scores 76.1 in BFCL, demonstrating strong tool use abilities.

ii. Nexus score of 38.5 is better than some models but significantly lower than LLaMA 3.1 70B (56.7).

iii. ZeroSCROLLS/QuALITY (81.0), InfiniteBench (65.1), and NIH/Multi-needle (98.8) scores indicate robust long-context understanding.

f) Multilingual Capabilities:

i. ConverseAI scores 68.9 in Multilingual MGSM (0-shot), outperforming Mistral 7B Instruct (29.9) but behind LLaMA 3.1 70B (86.9).

2) Strengths and Limitations

a) Strengths

i. Balanced Performance: ConverseAI performs well across multiple domains, especially in coding, general knowledge, and mathematical reasoning.

ii. Strong Long-Context Handling: With high scores in ZeroSCROLLS/QuALITY and NIH/Multi-needle, it effectively manages extended interactions.

iii. Competitive Reasoning & Tool Use: Scores in ARC Challenge and BFCL indicate solid problem-solving capabilities.

iv. Multilingual Proficiency: Above-average multilingual capabilities compared to

models like Mistral 7B and GPT-3.5 Turbo.

b) Limitations

i. Lower Performance in Advanced Reasoning: GPOQA scores suggest room for improvement in complex reasoning and commonsense understanding.

ii. Mathematical Challenges: Although strong in GSM8K, ConverseAI struggles in MATH (0-shot, CoT).

iii. Tool Use and Nexus Performance: While BFCL is strong, Nexus scores are relatively low, suggesting scope for better external tool integration.

3) Possible Improvements & Future Scope

a) Enhanced Fine-Tuning for Advanced Reasoning: Improving GPOQA and MATH performance through further fine-tuning on heavy-reasoning datasets.

b) Wider Context Window Expansion: Employing efficient memory management strategies to improve the understanding of wider contexts.

c) Improved External Tool Integration: Improved Nexus interactions to offer more external tool flexibility.

d) Multilingual Efficacy Optimization: Augmenting non-English training sets to improve performance on multi-language tasks.

e) Hardware Optimization: Investigating GPU acceleration and quantization for rapid inference without loss of accuracy.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] K. Eves and J. Valasek, “Adaptive control for singularly perturbed systems examples,” Code Ocean, Aug. 2023. [Online]. Available: https://codeocean.com/capsule/4989235/tree

[9] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114

[10] S. Liu, “Wi-Fi Energy Detection Testbed (12MTC),” 2023, gitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC