iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78760

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

A Multi-Purpose Text Generation Framework Using LLM: OWN-GPT

Kinal Patel[1], Rauki Yadav[2], Payal Naykawala[3], Bhumi Torani[4]
[1,2]*Department of Artificial Intelligence, Bhagwan Mahavir College of Engineering & Technology, BMU, Surat, India*
[3]*Department of Computer Engineering, Bhagwan Mahavir Polytechnic, BMU, Surat, India*
[4]*Department of Information Technology, Tapi Diploma Engineering College,Surat, India*

Abstract: ***OWN-GPT is a 2.3-billion-parameter transformer-based Large Language Model (LLM) developed as a full-stack AI system featuring a React.js frontend and a Python FastAPI backend. The system provides conversational text generation, image-based question answering using OCR and visual feature extraction, and PDF document question answering through Retrieval-Augmented Generation (RAG). The backend incorporates PyTorch, HuggingFace Transformers, PyMuPDF, LangChain, FAISS, sentence-transformers, OpenCV, Tesseract, and ViT-Base/16 for multimodal processing. Trained on an 81 GB multi-domain corpus (~40 billion tokens), OWN-GPT achieves a perplexity of 7.8, BLEU score of 0.68, image module CER of 6.8%, and RAG module faithfulness and relevancy scores of 0.81 and 0.79. Designed for local deployment in Indian academic institutions, the system ensures data privacy and zero per-query operational cost. A user study reports a System Usability Scale (SUS) score of 77.1, confirming strong usability and user satisfaction.***

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated exceptional performance in natural language understanding, text generation, multimodal reasoning, and domain-specific knowledge tasks. However, most state-of-the-art LLMs operate through cloud-based services, leading to concerns related to data privacy, high operational cost, and dependency on external infrastructure. These limitations are particularly significant for academic institutions in India, where sensitive student data and research documents must remain on-premise.

To address these challenges, OWN-GPT has been developed as a locally deployable, full-stack multimodal LLM system. Unlike cloud-based models, OWN-GPT operates entirely within an institution's local server environment, removing per-query costs and ensuring complete control over data. The system integrates conversational AI, image question answering using OCR and transformer-based visual encoders, and PDF question answering through Retrieval-Augmented Generation (RAG).

The architecture combines FastAPI-based backend inference, a React.js single-page application (SPA) frontend, and multiple specialized machine learning components for multimodal understanding. This paper presents the system design, training methodology, implementation details, evaluation metrics, and practical usability outcomes of OWN-GPT.

II. LITERATURE REVIEW

Recent advancements in transformer-based architectures have significantly accelerated the development of scalable and high-performance Large Language Models (LLMs). Foundational research such as *Attention Is All You Need* introduced the transformer architecture, which replaced recurrence and convolution with self-attention, enabling parallelism and improved contextual understanding [8]. This breakthrough laid the groundwork for subsequent models including BERT [11], T5 [12], GPT-2 [13], GPT-3 [7], and later open-source frameworks such as LLaMA 2 [9] and Pythia [10]. These models demonstrated exceptional capabilities in text generation, reasoning, and multilingual understanding, thereby setting new benchmarks for natural language processing. Comprehensive surveys on transformers and LLMs highlight the evolution from encoder-based architectures like BERT to decoder-focused models such as GPT-series, along with their increasing relevance in real-world applications [1], [3], [4], [6].

Parallel to text-only LLM improvements, multimodal processing has advanced with the integration of optical character recognition (OCR) and visual transformers. The evolution of OCR—from classical CRNN-based methods to modern systems like Tesseract—has significantly enhanced text extraction accuracy across complex image formats [5], [14]. Additionally, Vision Transformers (ViT) have enabled scalable representation learning for images, supporting multimodal tasks that combine textual and visual reasoning [4]. Studies on multimodal LLMs emphasize their role in improving accessibility, educational tools, and intelligent automation systems [14], [16], [17], [18].

©IJRASET: All Rights are Reserved

Despite these advancements, existing literature identifies several limitations in current LLM implementations. First, most high-performance models require cloud-based deployment, which raises significant privacy concerns for institutions handling sensitive academic, administrative, or student data. Research stresses the lack of locally deployable, privacy-preserving LLM frameworks tailored for institutions with regulatory or confidentiality constraints [2]. Second, commercial LLM services incur substantial API and operational costs, making them financially unsustainable for continuous high-volume usage in educational or research environments [1], [14]. Third, although multimodal models exist, their integration in open-source, institution-level systems remain limited. Many available models either lack OCR capabilities, fail to integrate document retrieval mechanisms, or do not provide end-to-end multimodal processing within a unified architecture [3], [4], [10].

OWN-GPT addresses these gaps by offering a fully on-premise, multimodal LLM system that supports conversational text generation, image-based question answering, and PDF-based retrieval-augmented generation. By eliminating cloud dependencies, the system ensures complete data privacy while avoiding per-query commercial costs. Its integrated architecture—combining transformer-based inference, OCR extraction, vision-transformer processing, and FAISS-driven RAG—directly responds to the shortcomings highlighted in existing research and provides a practical, deployable solution for academic institutions.

III. METHODOLOGY

The methodology for developing OWN-GPT consists of five major components: dataset preparation, model training, PDF RAG pipeline, image question-answering pipeline, and full-stack system design. The model is trained on an 81 GB multi-domain corpus comprising 25% Wikipedia, 30% Common Crawl, 15% BooksCorpus, 15% Q&A datasets, and 15% technical papers. Before training, the dataset undergoes extensive preprocessing, including tokenization, deduplication, sentence segmentation, and domain balancing to ensure high-quality and consistent input data.

For model training, OWN-GPT—a 2.3-billion-parameter transformer—is optimized using the AdamW optimizer across 15 training epochs with FP16 mixed precision. A cosine decay scheduler controls the learning rate throughout training, and the model is exposed to approximately 40 billion tokens. This training strategy ensures efficient convergence while maintaining stability across large-scale training cycles.

The system also features a PDF Retrieval-Augmented Generation (RAG) pipeline. In this module, PyMuPDF first extracts raw text from PDF documents, after which LangChain segments the text into manageable chunks suitable for retrieval. Sentence-transformers convert these chunks into vector embeddings, and FAISS performs fast similarity search to retrieve the most relevant segments. The LLM then generates answers grounded in the retrieved context, ensuring high factual accuracy.

Similarly, the Image Question Answering module integrates multiple computer vision components. OpenCV handles initial image preprocessing, while Tesseract v5 performs OCR to extract text. Visual features are generated using the ViT-Base/16 model, and the LLM fuses OCR text and visual embeddings to produce context-aware answers for image-based queries.

Finally, the system's full-stack architecture includes a React.js 18 frontend styled with Tailwind CSS and powered by Axios and React Router for API communication. The backend, built using FastAPI 0.110 and Python 3.11, handles all inference tasks using PyTorch 2.2 and HuggingFace Transformers. This combination results in a scalable, responsive, and production-ready multimodal AI system. Additionally, the architecture supports modular integration, allowing new features and models to be incorporated with minimal changes.

It is designed for efficient parallel processing to handle multiple user requests simultaneously. The system also emphasizes low-latency inference for real-time interactions.

Furthermore, it can be extended with cloud-based scaling solutions for high availability and performance. Continuous monitoring and logging mechanisms are incorporated to ensure reliability and maintain system health.

IV. SYSTEM ARCHITECTURE

OWN-GPT follows a modular architecture with three core subsystems:

A. Language Model Inference Module

The Language Model Inference Module forms the core of OWN-GPT and is responsible for processing conversational inputs, performing reasoning tasks, and generating coherent, context-aware responses. It interprets user queries, analyzes intent, and produces fluent text outputs across multiple domains, enabling natural and meaningful human–AI interaction.

B. Image Processing Module

The Image Processing Module enhances OWN-GPT with the ability to understand and analyze visual content. It incorporates OCR techniques to extract textual information from images and utilizes a Vision Transformer (ViT) model to capture high-level visual features. By combining OCR output with visual embeddings, this module enables the system to answer questions related to diagrams, documents, photographs, and other image-based inputs.

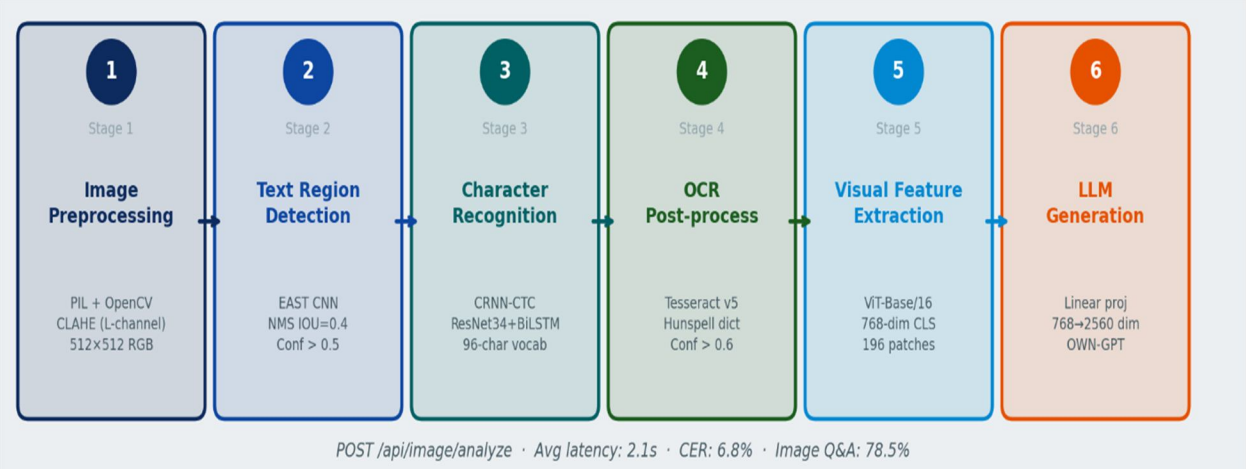


Figure 1: Image Processing Module Six-Stage Pipeline

C. PDF RAG Module

The PDF RAG Module enables document-level question answering by integrating retrieval-augmented generation techniques. It extracts text from PDF documents, breaks it into manageable chunks, encodes them into vector embeddings, and retrieves the most relevant segments using FAISS. The language model then generates responses grounded in the retrieved context, ensuring accurate and evidence-based answers for PDF-based queries.

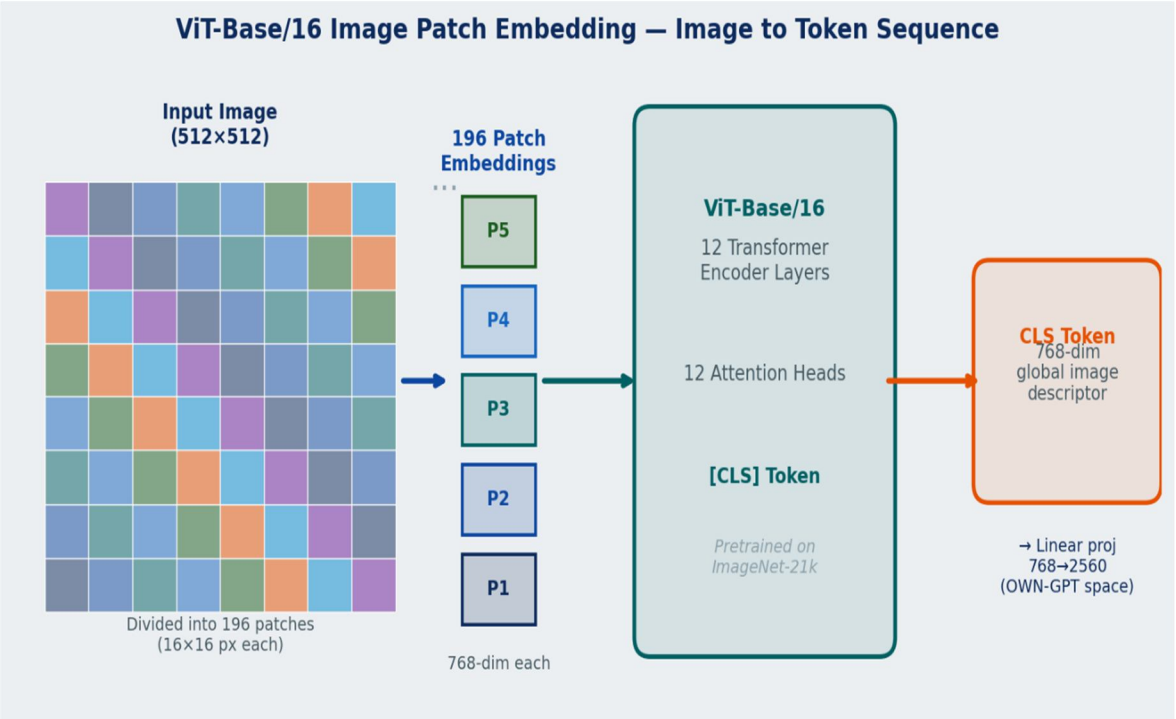


Fig 2: PDF Processing and RAG Module Pipeline

Each module communicates with the backend API, while the React-based frontend ensures seamless user interaction.

V. RESULTS AND PERFORMANCE EVALUATION

The performance of OWN-GPT was evaluated across its language, image, and document-retrieval components, along with a usability assessment to measure user acceptance. The Language Model Inference Module demonstrated strong generative quality, achieving a perplexity of 7.8 and a BLEU score of 0.68.

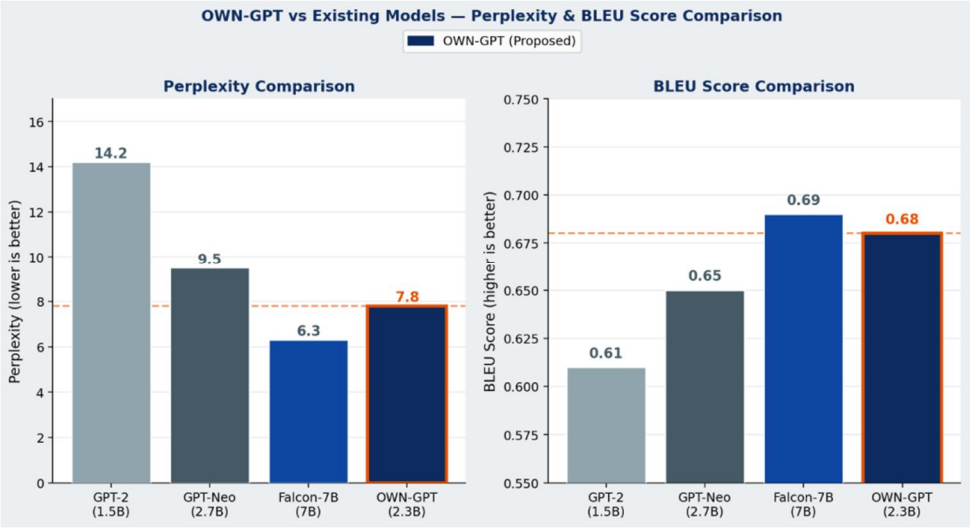


Fig 3: Perplexity and BLEU Score comparison bar charts

Human evaluators rated the model's coherence at 4.6/5 and fluency at 4.7/5, indicating consistent and contextually accurate text generation. In the Image Processing Module, the system recorded a Character Error Rate of 6.8% and achieved an image-based question answering accuracy of 78.5%, reflecting its capability to integrate OCR output with visual feature representations.

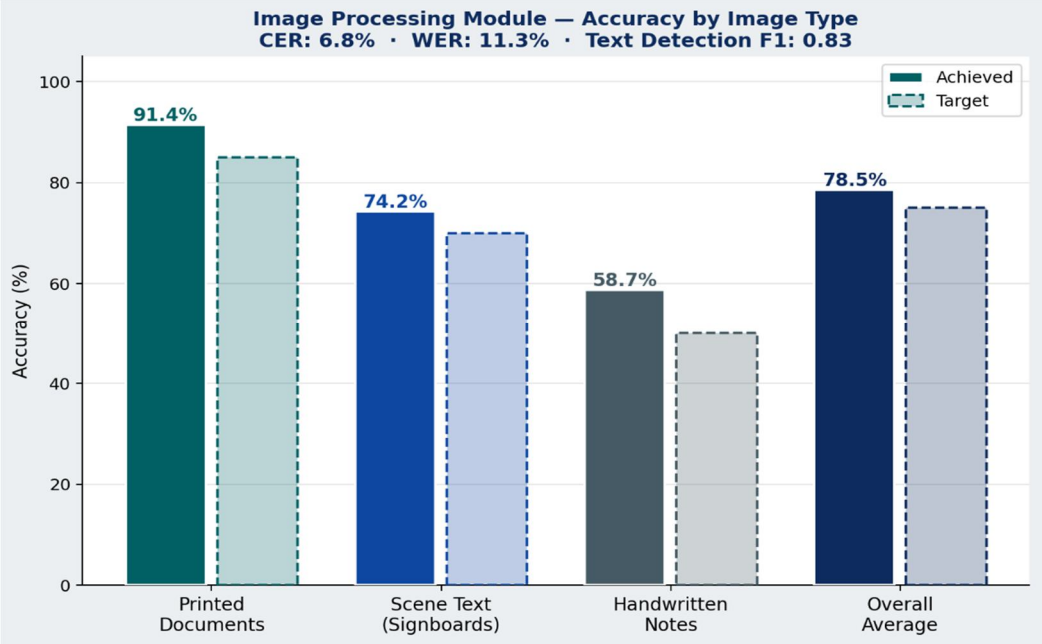


Fig 4: Image Module OCR Accuracy by Image Type

The PDF RAG Module also exhibited effective grounded reasoning, with a RAGAS Faithfulness score of 0.81 and an Answer Relevancy score of 0.79, demonstrating reliable context-based retrieval. A user study conducted at BMU Surat further validated the system's practicality, reporting a System Usability Scale score of 77.1 and an overall satisfaction level of 4.2/5. These combined results confirm that OWN-GPT provides accurate, efficient, and user-friendly multimodal AI performance suitable for academic environments.

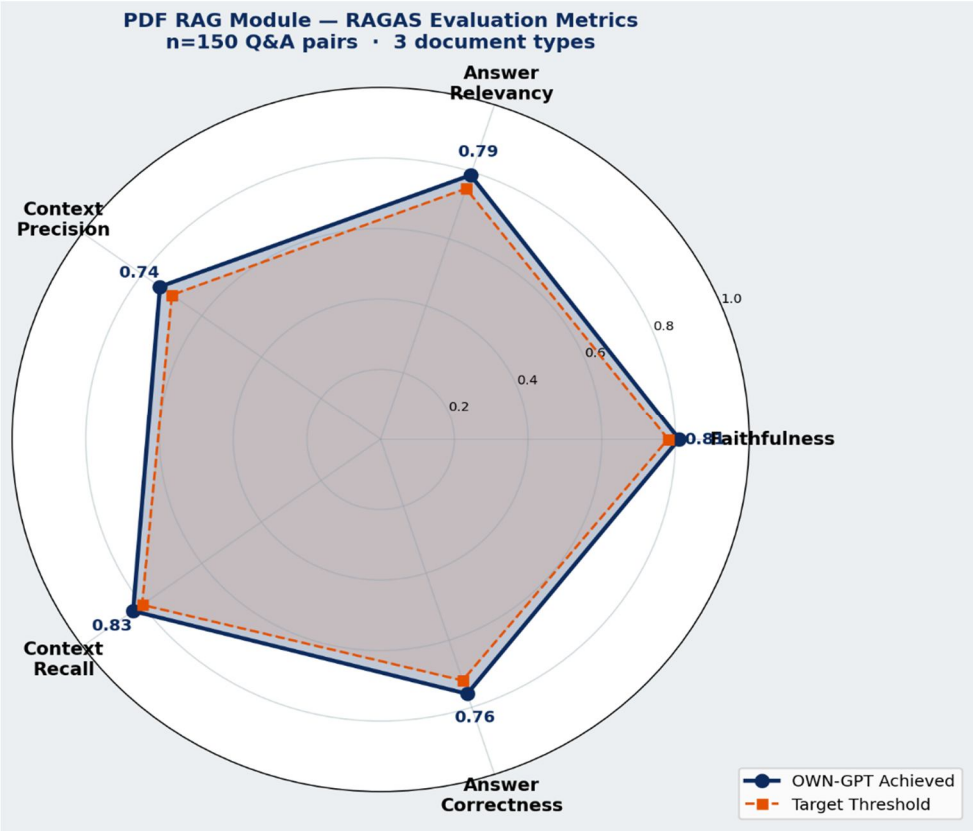


Fig 5: PDF RAG RAGAS Metrics Radar Chart

VI. DISCUSSION

The evaluation indicates that OWN-GPT provides performance comparable to mid-size LLMs while offering unique benefits such as on-premise deployment and multimodal capabilities. Its modular backend allows easy customization for academic courses, research assistance, and administrative automation. The PDF RAG module particularly enhances academic utility by enabling question answering directly from research documents.

VII. CONCLUSION

OWN-GPT demonstrates that a fully local, multimodal LLM system can be successfully developed and deployed within academic institutions. Its strong performance across text generation, OCR-based image understanding, and PDF-RAG-based document analysis makes it a valuable tool for teaching, research, and administration. The system's cost-free operation post-deployment addresses major barriers faced by educational organizations.

VIII. FUTURE SCOPE

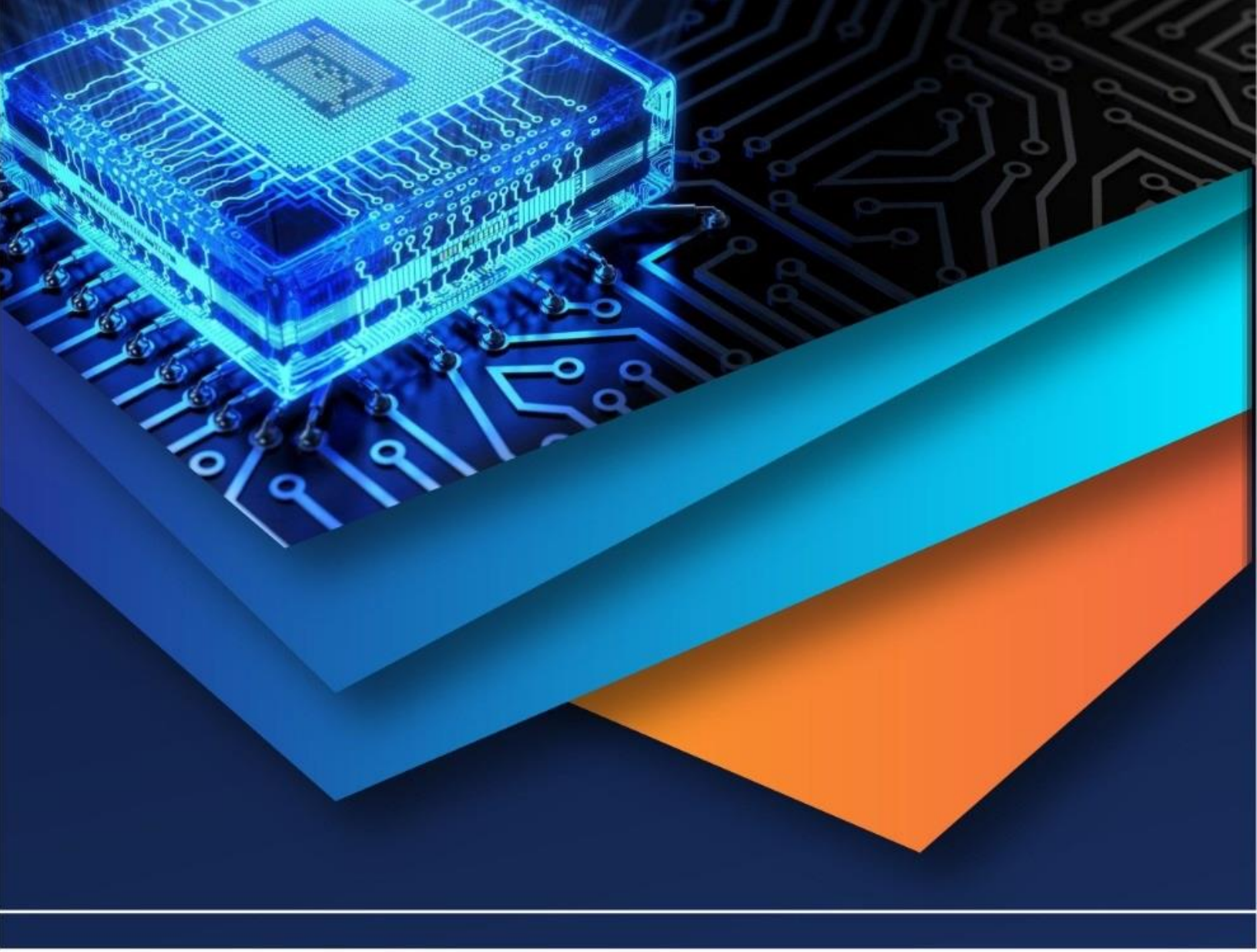
Future work will focus on scaling the model to 6–7 billion parameters to improve performance, while also fine-tuning it for specific academic domains. Additional enhancements include integrating speech-to-text and text-to-speech capabilities for better accessibility, enabling real-time multi-user load balancing to support concurrent usage, and optimizing the system for on-device mobile deployment. Furthermore, integration with institutional LMS platforms will be pursued to streamline adoption in educational environments.

REFERENCES

[1] V. Visweswaraiah, "Everyday AI: Real-World Applications of Transformer-Based Large Language Models," IJCTT, vol. 73, no. 9, pp. 19-27, 2025. [Online]: https://ijcttjournal.org
[2] A. Singh et al., "Research Gaps in Developing Fair and Inclusive LLMs for India's Agriculture Sector," IRJET, vol. 12, no. 11, pp. 550-556, 2025.
[3] M. A. Khan, R. Sharma, S. Patel, "Large Language Models: An Overview of Foundational Architectures, Recent Trends, and a New Taxonomy," Discover Applied Sciences, Springer Nature, doi:10.1007/s42452-025-07668-w, 2025.

[4] G. Yenduri et al., "GPT (Generative Pre-Trained Transformer) — A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," IEEE Access, vol. 12, pp. 54608-54649, doi:10.1109/ACCESS.2024.3387368, 2024.
[5] S. Suresh, M. Chandrika, "Large Language Model (LLM): An AI Model for Pattern Recognition," IJERT, 2023.
[6] A. Gupta, R. Rao, S. Mehta, "Advancements in Transformer Architectures for Large Language Models: From BERT to GPT-3 and Beyond," IRJMETS, vol. 6, no. 4, 2024.
[7] T. B. Brown et al., "Language Models are Few-Shot Learners," Advances in NeurIPS, vol. 33, pp. 1877-1901, 2020.
[8] A. Vaswani et al., "Attention Is All You Need," NeurIPS, vol. 30, arXiv:1706.03762, 2017.
[9] H. Touvron et al., "LLaMA 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, 2023.
[10] S. Biderman et al., "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling," ICML 2023, arXiv:2304.01373.
[11] J. Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT 2019, pp. 4171-4186, doi:10.18653/v1/N19-1423.
[12] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5)," JMLR, vol. 21, no. 140, 2020.
[13] A. Radford et al., "Language Models are Unsupervised Multitask Learners (GPT-2)," OpenAI Blog, 2019.
[14] Z. T. Hamad et al., "ChatGPT's Impact on Education and Healthcare," IEEE Access, vol. 12, pp. 114858-114876, doi:10.1109/ACCESS.2024.3437374, 2024.
[15] P. Liu et al., "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP," ACM CSUR, vol. 55, no. 9, doi:10.1145/3560815, 2023.
[16] M. Ansari et al., "Intelligent Chatbot," IJERT, vol. 10, no. 3, 2021.
[17] A. Binu et al., "Chatbot Using Artificial Intelligence," IJERT, vol. 12, no. 6, 2023.
[18] M. Senthil Kumar et al., "An Automated Chatbot for an Educational Institution using NLP," IJCRT, vol. 10, no. 5, pp. 142-150, 2022.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)