© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1719389

Interactive AI Conversation Using Small Language Model (SLM) Chatbot

T PRIYA[1], I ESTHER PRAISHE[2]
[1]*Assistant Professor, Computer Science and Engineering, AVS College of Technology, Attur Main Rd, Chinnagoundapuram, Salem.*
[2]*Student (M.E Computer Science and Engineering), AVS College of Technology, Attur Main Rd, Chinnagoundapuram, Salem.*

Abstract- ***Conversational Artificial Intelligence (AI) has transformed human–computer interaction by enabling intelligent, context-aware, and natural communication. However, most chatbot systems rely on Large Language Models (LLMs), which demand significant computational resources, memory, and cloud infrastructure. This paper presents an Interactive AI Conversation System based on a Small Language Model (SLM) chatbot that provides efficient, low-latency, and cost-effective conversational capabilities. The proposed system employs natural language processing (NLP), transformer-based SLM architecture, intent recognition, contextual memory, and response generation to deliver interactive conversations while operating with significantly lower computational requirements. The chatbot is designed for educational assistance, customer support, healthcare guidance, and enterprise applications. Experimental evaluation demonstrates reduced inference latency, lower memory consumption, and competitive conversational quality, making the proposed system suitable for deployment on edge devices and resource-constrained environments. Small Language Models are increasingly attractive because they provide lower latency and lower resource usage while remaining effective for many specialized conversational tasks.***

Keywords— Small Language Model, Chatbot, Artificial Intelligence, Natural Language Processing, Transformer, Conversational AI, Edge Computing.

I. INTRODUCTION

Artificial Intelligence has revolutionized human-computer interaction through conversational agents capable of understanding and generating natural language. Modern chatbots are widely deployed in education, healthcare, banking, e-commerce, and customer support. Although Large Language Models (LLMs) provide excellent conversational capabilities, they require extensive computational resources, high memory, and cloud-based infrastructure.

Small Language Models (SLMs) provide an efficient alternative by offering faster inference, lower memory consumption, and reduced operational cost. These characteristics make SLMs well suited for mobile devices, embedded systems, and enterprise applications where privacy, latency, or offline capability is important.

This paper proposes an Interactive AI Conversation chatbot using an SLM architecture that supports intelligent dialogue while maintaining computational efficiency.

II. RELATED WORK

Recent developments in conversational AI have focused on transformer-based language models for dialogue generation. Large Language Models demonstrate remarkable conversational ability but remain computationally expensive.

Recent studies indicate that SLMs can achieve competitive performance in domain-specific applications while reducing inference time and hardware requirements. Research has also shown that fine-tuned SLMs can outperform much larger models on narrowly defined application tasks.

III. PROPOSED SYSTEM

The proposed chatbot consists of the following modules:

1. User Interface

2. Input Processing
3. Natural Language Processing
4. Small Language Model
5. Context Memory
6. Response Generation
7. Output Display

Working Procedure

- User enters a text query.
- Input is preprocessed using tokenization and normalization.
- The SLM extracts semantic information using transformer attention.
- Context memory maintains conversation history.
- The response generation module produces the most relevant reply.
- The generated response is displayed to the user.
- The architecture minimizes memory usage while maintaining conversational quality.

IV. SYSTEM ARCHITECTURE

The system architecture contains four major layers:

- Presentation Layer
- NLP Processing Layer
- Small Language Model Inference Layer
- Response Generation Layer

The transformer-based SLM performs language understanding using self-attention mechanisms while requiring significantly fewer parameters than conventional LLMs, enabling efficient deployment on edge hardware.

V. EXPERIMENTAL RESULTS

The chatbot was implemented using Python together with a transformer-based SLM and evaluated using standard conversational benchmarks.

Performance metrics included:

- Response Accuracy
- Response Time
- Precision
- Recall
- User Satisfaction
- Memory Utilization

Experimental evaluation showed that the SLM chatbot achieved fast response generation with substantially lower memory usage than LLM-based systems while maintaining satisfactory conversational quality for domain-specific interactions.

VI. CONCLUSION

This paper presented an Interactive AI Conversation chatbot based on a Small Language Model. The proposed framework demonstrates that efficient conversational AI can be achieved without the computational demands of very large language models. The architecture offers reduced latency, lower resource consumption, and scalable deployment across edge devices, mobile platforms, and enterprise environments.

Future work includes multilingual conversation support, speech recognition integration, emotion-aware dialogue generation, and multimodal interaction using text, speech, and images.

REFERENCES

[1] A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[3] T. B. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.

[4] A. Radford et al., “Improving Language Understanding by Generative Pre-Training,” OpenAI, 2018.

[5] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv preprint arXiv:1907.11692, 2019.

[6] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” in Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[7] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[8] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. M. Turing, “Computing Machinery and Intelligence,” Mind, vol. 59, no. 236, pp. 433–460, 1950.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in Advances in Neural Information Processing Systems (NeurIPS), 2013, pp. 3111–3119.

[11] M. Chen et al., “Evaluating Large Language Models Trained on Code,” arXiv preprint arXiv:2107.03374, 2021.

[12] S. Min et al., “Recent Advances in Large Language Models,” ACM Computing Surveys, vol. 56, no. 9, 2024.

[13] M. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” arXiv preprint arXiv:2108.07258, 2021.

[14] Z. Li et al., “Small Language Models: Survey, Measurements, and Insights,” arXiv preprint, 2024.

[15] J. Wei et al., “Emergent Abilities of Large Language Models,” Transactions on Machine Learning Research (TMLR), 2022.