Hybrid Documents RAG based Chatbot

Maharun Afroz, *2211023642,* Sanjida Amin Nadia, *2122319642,* Tasnia Hossain, *2121480642,*
and Tahmina Mozumdar, *2111482642*

I. Literature Review

RAG-based chatbots combine offline and online modes to enhance security and data privacy. Offline models extract information from PDFs using embeddings, while online models fetch real-time data via APIs. This hybrid approach ensures privacy with offline access and up-to-date info through online search, though challenges in response accuracy and optimization persist.

Unimib Assistant: This chatbot, crafted for university students, showcased the benefits of RAG in accessing pertinent academic information. Nevertheless, it also uncovered difficulties such as inaccuracies, absence of relevant data, and broken links. These observations highlight the necessity for effective prompt engineering and user-centered design to enhance the usability and reliability of chatbots [1].

AWAITS Project: This research aimed at boosting the dependability of educational AI chatbots by incorporating RAG to mitigate misinformation and improve response precision. Comparative analyses between chatbots utilizing RAG and those that do not indicate that real-time retrieval markedly elevated the relevance and trustworthiness of the responses generated, especially in the context of academic writing assistance [2].

II. Work Progress

We have interacted with several SLMs and LLMs through LangChain and also tried out custom prompts.

Langchain works as an orchestrator, making it easier to interact with models of different versions and parameters. The models can be given specific instructions about their behavior. Also, temperature for different models where a higher value is more creative, lower is more coherent. We can also decide whether a model's response would be verbose or quiet.

Some of the models that we explored are:

Model Name	Parameter	Size	Use Cases
qwen	0.5b	394 MB	Lightweight, fast language model for quick responses and small-scale applications.
gemma	2b	1.7 GB	Suitable for more complex natural language processing tasks with moderate resource requirements.
gemma	7b	5.0 GB	Ideal for in-depth text generation, summarization, and understanding with higher accuracy.
deepseek-r1	1.5b	1.1 GB	Useful for research and experimentation with small to medium-sized datasets.
deepseek-r1	8b	4.9 GB	High-performance tasks involving natural language understanding, multi-turn conversations, and detailed text analysis.
llama3.1	8b	4.7 GB	General-purpose model for a variety of applications including text generation, question answering, and summarization.
llama2-uncensored	7b	3.8 GB	Optimized for applications requiring uncensored, diverse language generation or open-domain conversations.
bakllava	7b	4.7 GB	Focused on creative writing and generating content with diverse linguistic styles.
codellama	7b	3.8 GB	Specializes in code generation, programming-related tasks, and assisting with software development.

TABLE I: Model Parameters and Use Cases

References

[1] C. Antico et al., ”Unimib Assistant: Designing a Student-Friendly RAG-Based Chatbot for All Their Needs,” in *Proc. Italian Workshop Artif. Intell. Human-Machine Interact. (AIxHMI)*, 2024.

[2] K. Matar and Y. Mohammad, ”Improving the Reliability of Educational AI Chatbots Using Retrieval-Augmented Generation,” M.S. thesis, Linnaeus Univ., Sweden, 2024.