Generative AI Study Assistant using RAG

Team Name: Rahaf Kanaan, Shifaa Al-zu'bi, Thabet Zamari, Rafah Ali, Tasneem Alassaf.

Building the RAG-based chatbot was a valuable hands-on experience that connected theoretical concepts from the Generative AI course with practical implementation. One of the aspects that worked particularly well was integrating the document retrieval pipeline with the language model, which ensured that the chatbot's answers were grounded in the course material rather than relying on generic responses. The conversational interface also enhanced usability and made the system intuitive to interact with.

During development, several limitations were encountered, mainly related to model selection and computational constraints. Initial experiments with local open-source models such as *flan-t5-small* resulted in fast but often imprecise answers. Switching to *mistralai/Mistral-7B-Instruct-v0.2* improved answer quality but introduced significant latency, making the chatbot impractical to use. Eventually, *flan-t5-base* provided a reasonable balance between speed and accuracy, while API-based models like *GPT-4o* delivered the most coherent and detailed explanations.

Overall, the use of RAG significantly improved answer relevance by grounding responses in retrieved lecture content, reducing hallucinations and increasing trustworthiness. This project demonstrated how combining retrieval with generation can enhance both the reliability and educational value of AI-powered study assistants.

GoogleColab-Link:

https://colab.research.google.com/drive/1E-epswJmLOC-U95PgrE1jBdx-VIUvs2R?usp=sharing