

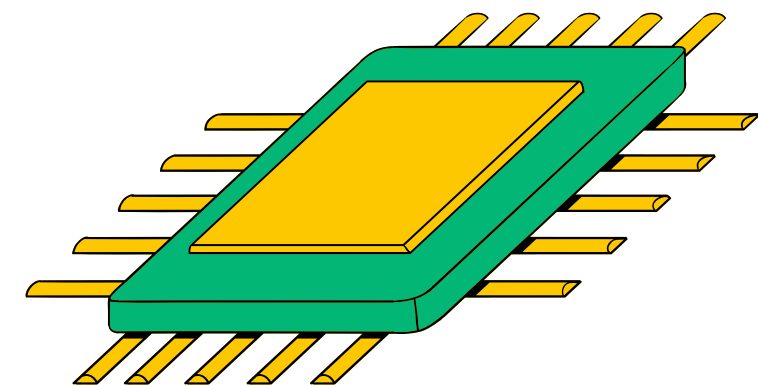
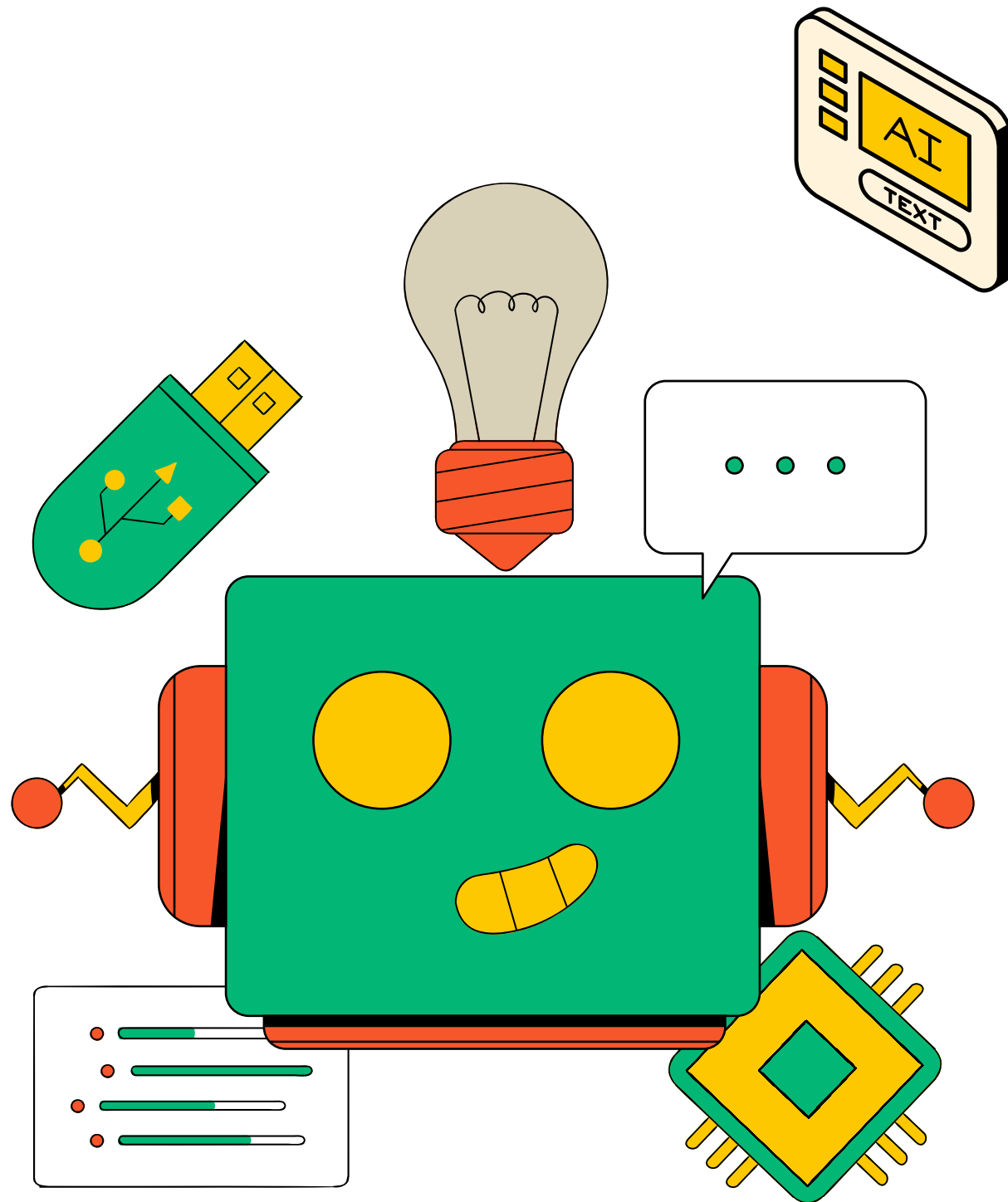


AI TEACHING ASSISTANT

PRESENTATION

PRESENTED BY:

GROUP NO 4



PRESENTATION OUTLINE

R

- Objectives
- Document Loading
- Document Splitting and Metadata
- Embedding and Vector Store
- Retriever and Similarity Search
- Prompt & LLM for QA
- RAG Working Flow
- Experiments and Results
- Limitations and Future Scope
- Q&A



OBJECTIVES OF THIS PROJECT

- Interactive guidance on course materials.
- NLP-driven educational tool for enhanced learning.
- Autonomous Answer Question system handling for course queries.
- Improved understanding, personalized assistance.
- Enhanced learning, and efficient support.



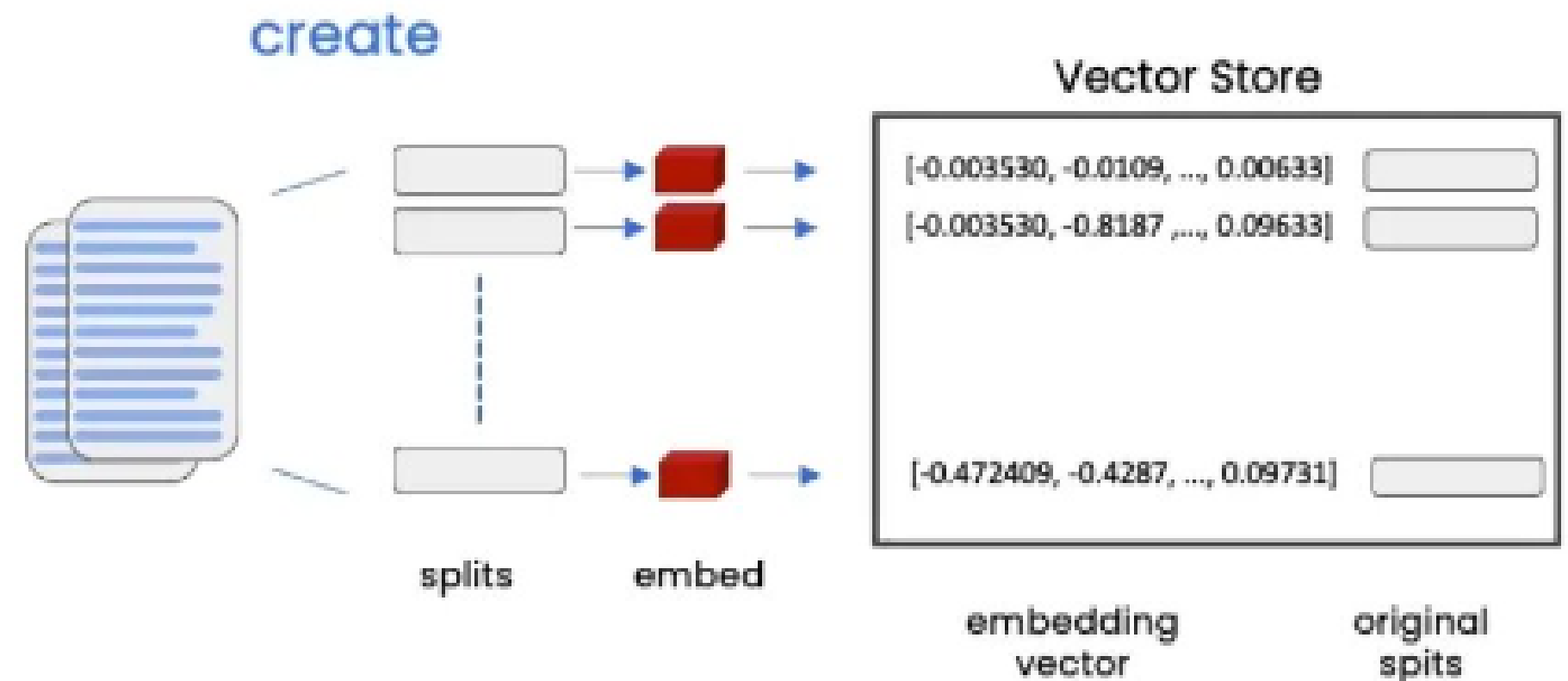
DOCUMENT LOADING AND TEXT EXTRACTION:

- Loaded all documents into Jupyter notebook from google drive using file ids.
- Extracted text as well as other metadata (File title, Last modified time, Source) of each pages using "pdfminer" library to build an index.
- Used "Recursive Text Splitter" of langchain to split extracted text into number of chunks in recursive order with overlapping. Overlapping allowed for to have some notion of consistency between 2 chunks.



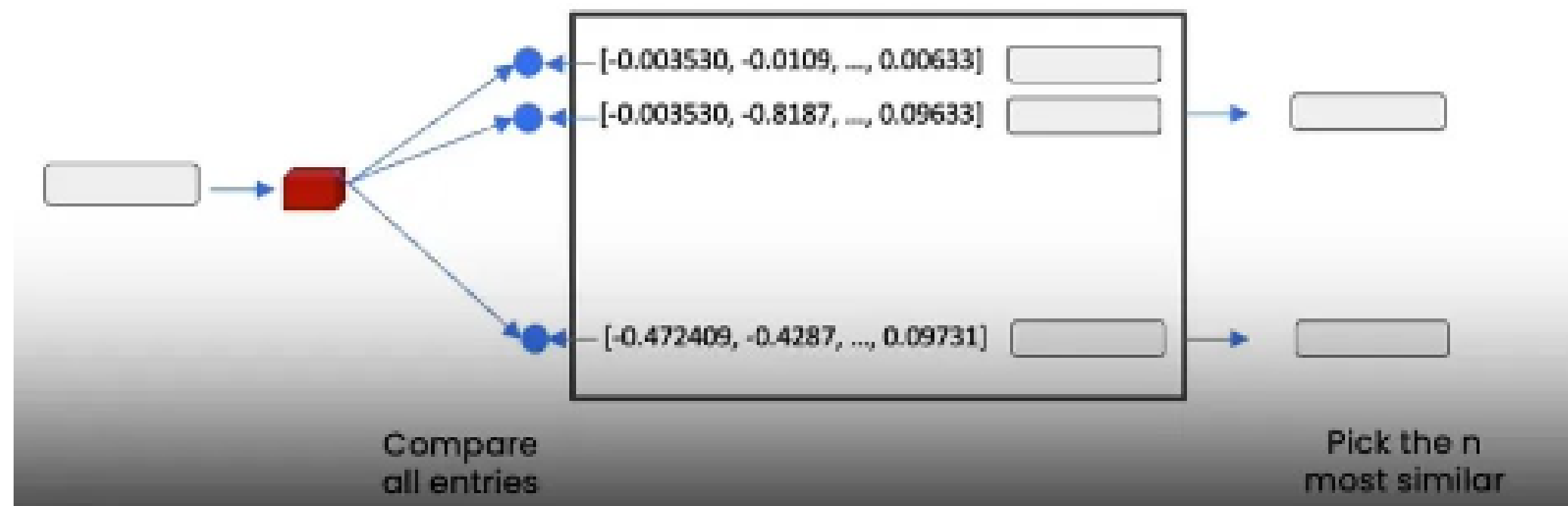
EMBEDDING AND VECTOR STORE:

- The Sentence Transformers library is utilized to generate dense vector representations (embeddings) of text chunks. These embeddings are computed using pre-trained transformer models, such as "all-MiniLM-L6-v2". Each chunk of text is encoded into a dense vector representation, capturing its semantic meaning.
- FAISS (Facebook AI Similarity Search) is a library for efficient similarity search and clustering of dense vectors. It is used here for building and querying a vector database using dense vector representation.



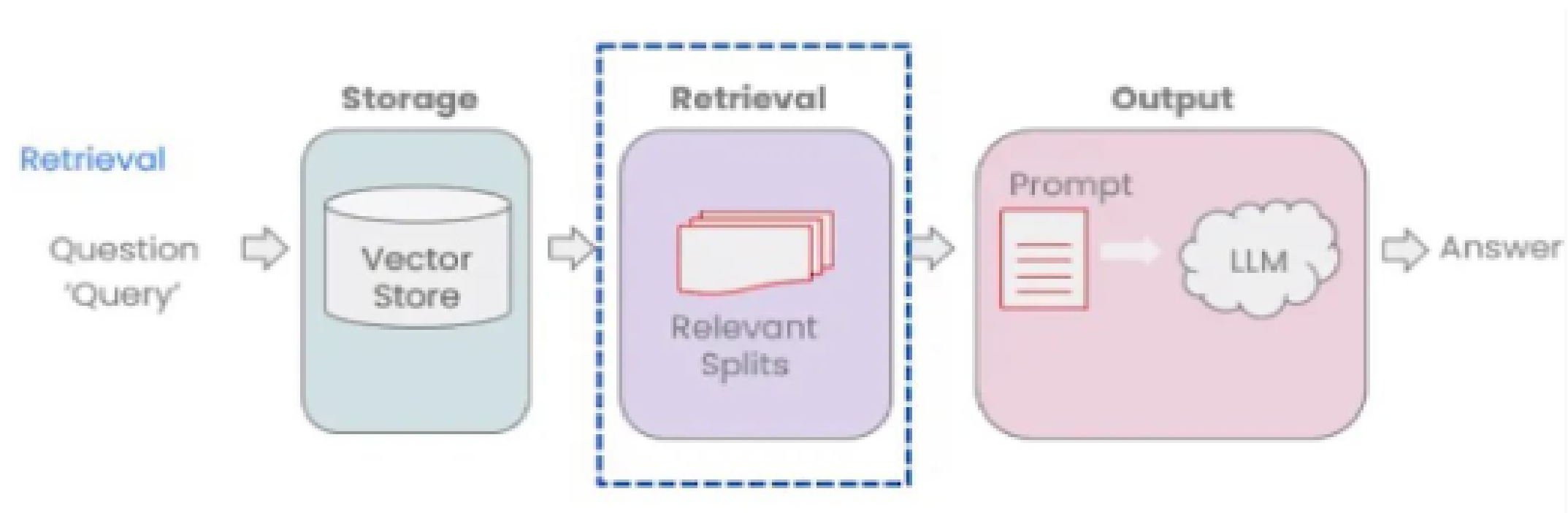
RETRIEVER AND SIMILARITY SEARCH:

- After setting up the vector database, a retriever is configured using `vectordb.as_retriever()`. This retriever likely uses the vector embeddings of documents to retrieve relevant context for a given user query efficiently.



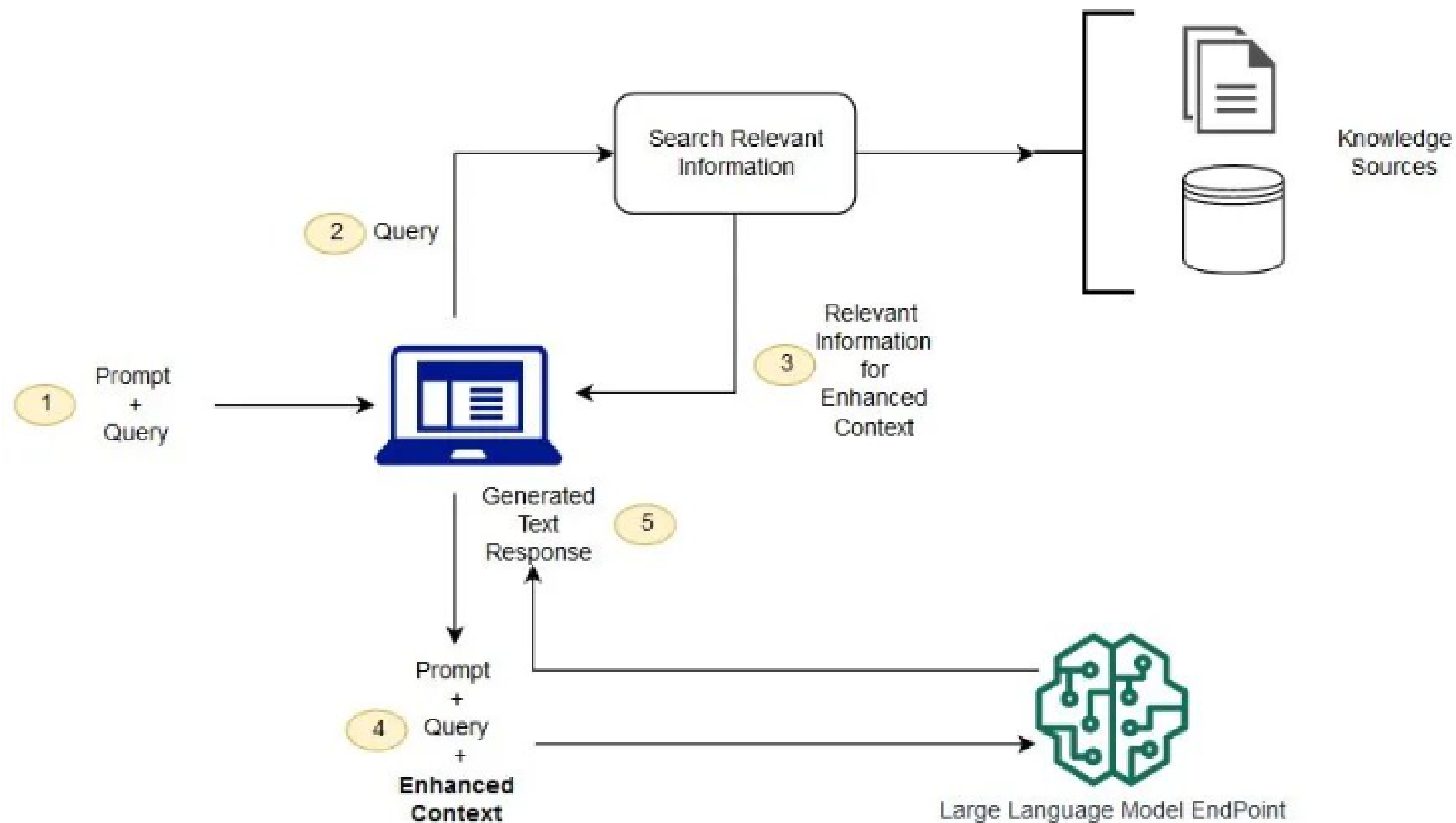
PROMPT & LLM FOR QA

- A prompt template is defined for generating responses in a conversational format. The template provides instructions for generating answers based on given contexts and user questions.



- A prompt template is defined for generating responses in a conversational format. The template provides instructions for generating answers based on given contexts and user questions.

RAG WORKING FLOW



TESTING SYSTEM OUTPUT

USER: HI THERE!

MAC AI ASSISTANT: HEY THERE! HOW CAN I HELP YOU TODAY?

CITATION: TITLE: ALL_LECTURE_PDFS\LECTURE09-MULTIMODAL MODELS.PDF PAGE 79 SOURCE: EMPTY_URL

USER: CAN YOU PROVIDE ME COURSE LOGISTICS IN BRIEF?

MAC AI ASSISTANT: OF COURSE! THE COURSE WILL RUN FROM WEDNESDAY 3:30PM-6:30PM. PLEASE REFER TO AVENUE TO LEARN FOR ADDITIONAL INFORMATION REGARDING COURSE WORK AND GRADING POLICY. THE ASSIGNMENTS ARE WORTH 20%, QUIZZES ARE WORTH 10%, MIDTERM IS WORTH 35%, AND THE FINAL GROUP PROJECT IS WORTH 34%. LATE DAYS WILL RESULT IN A DEDUCTION OF 1% OFF THE COURSE GRADE PER DAY LATE, AND ASSIGNMENTS WILL NOT BE ACCEPTED MORE THAN 3 DAYS LATE UNLESS GIVEN PERMISSION IN ADVANCE. COLLABORATION IS ALLOWED, BUT YOU MUST DOCUMENT HOW YOU COLLABORATED AND ACKNOWLEDGE ANY OUTSIDE HELP. PROJECT PROPOSALS WILL BE POSTED THIS WEEK, AND THE A2 IS DUE THIS SATURDAY BY 11:59PM. IF YOU ENCOUNTER ANY ISSUES WITH COLAB, PLEASE REACH OUT TO US FOR ASSISTANCE.

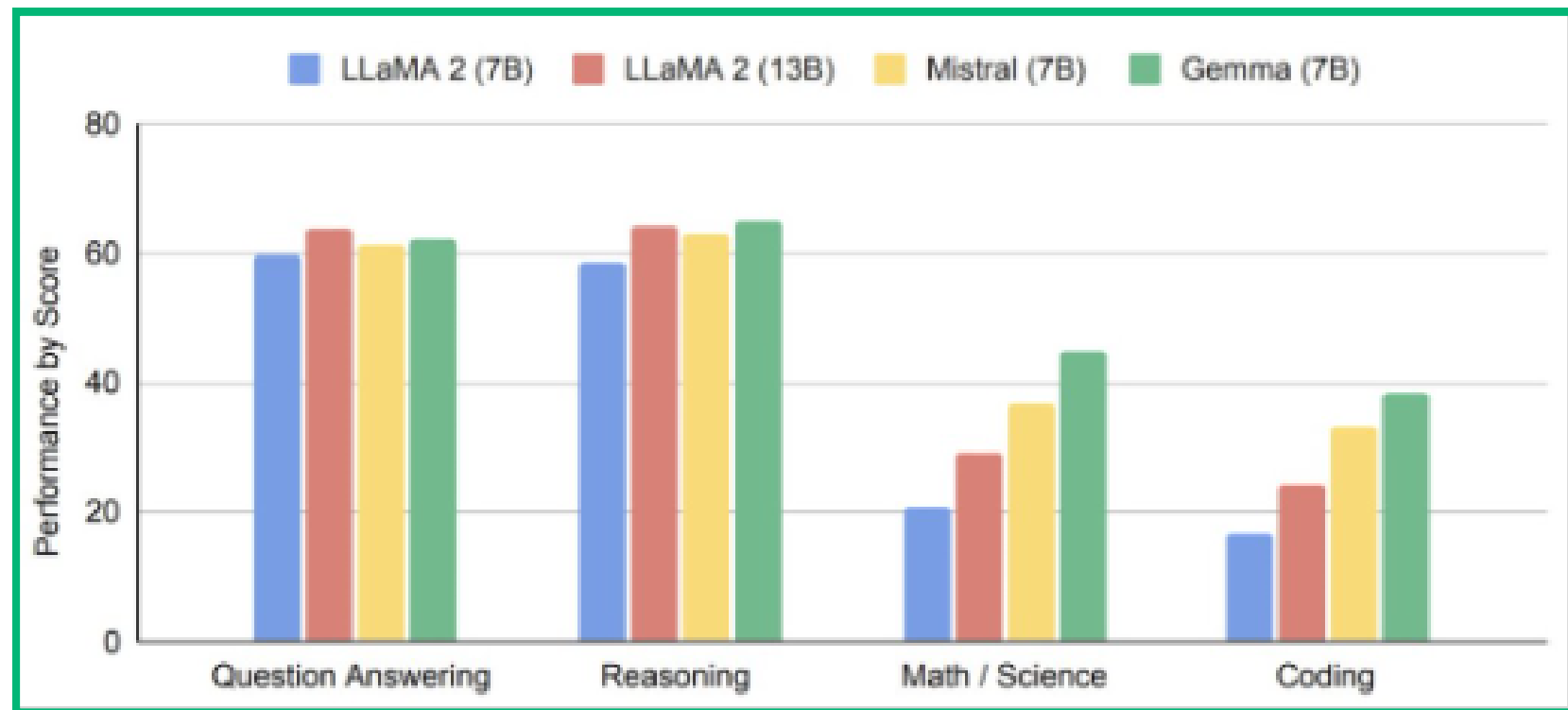
CITATION: TITLE: ALL_LECTURE_PDFS\LECTURE01-WORDVECS1.PDF PAGE 3 SOURCE: EMPTY_URL

USER: WHAT DO YOU MEAN BY RNN?

MAC AI ASSISTANT: RNN STANDS FOR RECURRENT NEURAL NETWORK, IT IS A TYPE OF NEURAL NETWORK ARCHITECTURE THAT IS PARTICULARLY USEFUL FOR PROCESSING SEQUENTIAL DATA SUCH AS TIME SERIES DATA OR NATURAL LANGUAGE TEXT.

CITATION: TITLE: ALL_LECTURE_PDFS\LECTURE03-RNNS AND LLMS.PDF PAGE 71 SOURCE: EMPTY_URL

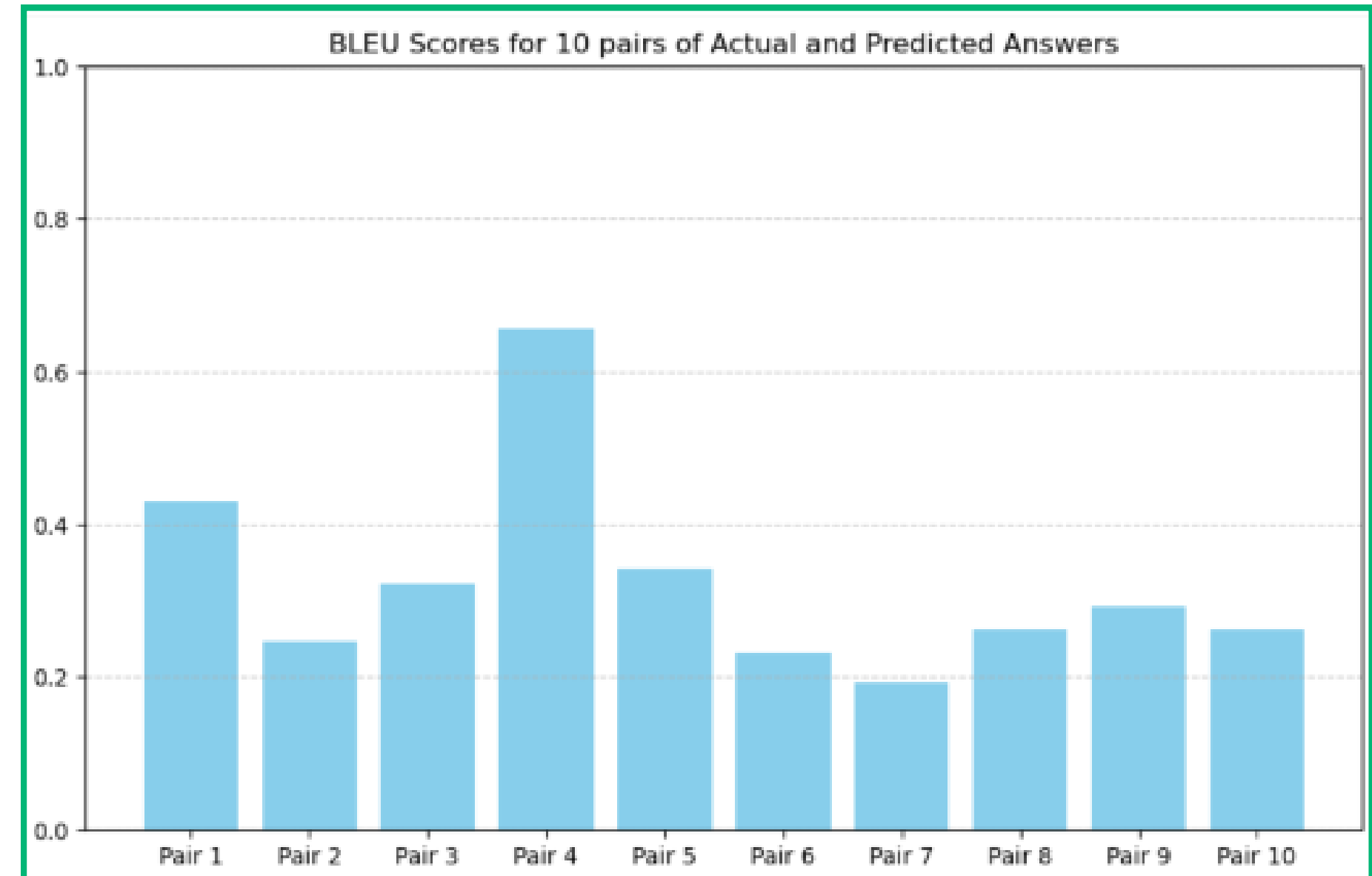
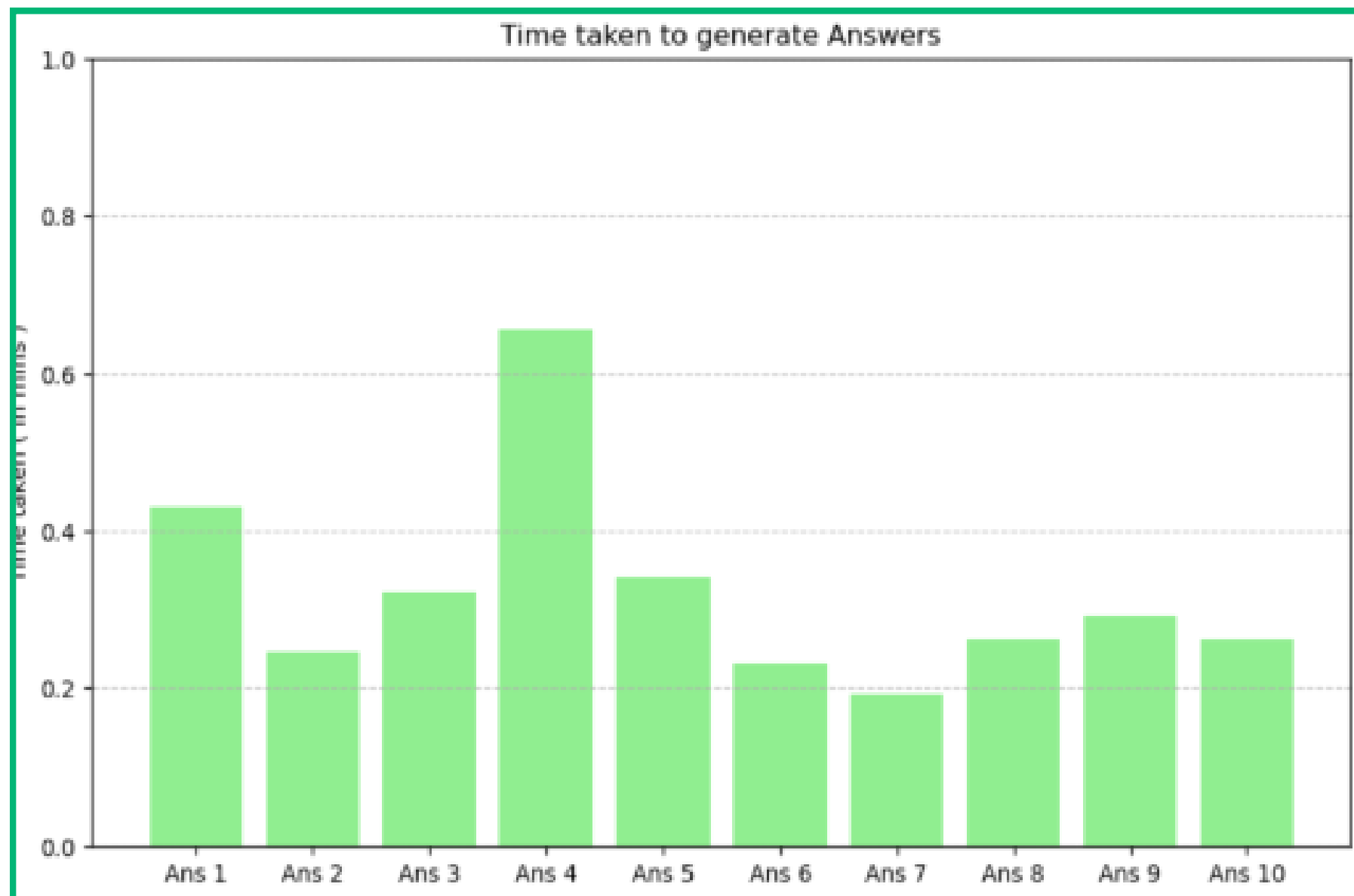
EXPERIMENTS AND RESULTS



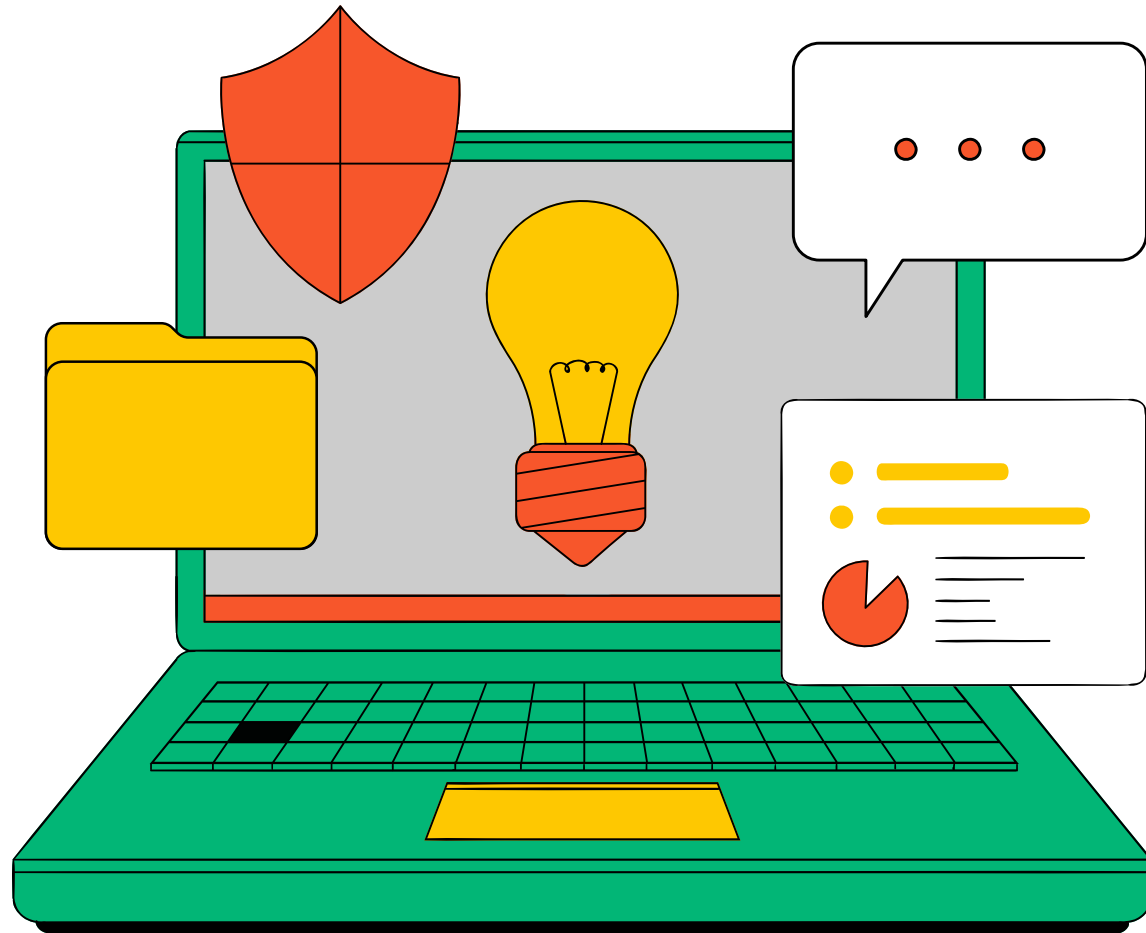
- Explored varied k values for optimal retrieval.
- Evaluated Gemma2-7b, Llama2-13b, and Llama2-7b.
- Llama2-7b outperformed other models consistently.
- Extracted 10 QA pairs from lecture PDF.
- Profiled BLEU scores and response generation time.



EXPERIMENTS AND RESULTS



LIMITATIONS & FUTURE SCOPE



- Boost search score with metadata for recent information.
- Show retrieved content with source URL("Source:" URL) to check language model's hallucinations.
- Include chat history(Buffer Memory) into prompt for interactive conversation experience.
- Store text and image embeddings for multi-model RAG.
- Improving AI Prompts for Better and Precise and Domain Specific Responses:



FRONTEND DESIGN 1





Menu:

Upload your PDF Files and Click on the Submit & Process Button

Drag and drop files here

Limit 200MB per file

Browse files

-  Yolo.pdf
5.1MB 
-  Attention.pdf
2.1MB 

Submit & Process

Done

Ask Questions | McMaster University

Ask a Question from the PDF Files



FRONTEND DESIGN 2

McMaster Instructor Panel:

Menu:

Upload your PDF Files and Click on the Submit & Process Button

Drag and drop files here

Limit 200MB per file

Browse files

Submit & Process

How It Works

Follow these simple steps to interact with the Mac AI Assistant:

1. **Upload Your Documents:** The system accepts multiple PDF files at once, analyzing the content to provide comprehensive insights.
2. **Ask a Question:** After processing the documents, ask any question related to the content of your uploaded documents for a precise answer.

McMaster University | AI Teaching Assistant 🧑🏫

Ask a Question



REFERENCES



- Meta AI. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. Retrieved March 21, 2024, from Meta AI.
- AWS. (2023). What is RAG? – Retrieval-Augmented Generation Explained. Amazon Web Services, Inc. Retrieved March 21, 2024, from AWS.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Retrieved from SBERT.net.
- Mahayana, D. Y. (2023, December 27). Training Your Own Dataset in Llama2 using RAG LangChain. Medium. Retrieved March 25, 2024, from Medium.



THANK YOU

