Gen AI Fusion - Challenge 3 (Academic Articles)

Team: insight.ai

Daichi Ishikawa
Josh Rayvika
Sivaram Mandava

Live UI Link: http://209.97.145.117:8501
GitHub Link: https://github.com/daichi6/llm-hackathon-insightai/tree/main

July 2024

Data Preprocessing

- By leveraging PyMuPDF and AWS Textract, we successfully extracted images and tables.
- Additionally, by extracting images with buffer to include image and table numbers, we ensured that they could be referenced later.

Image/Figure Extraction (PyMuPDF)

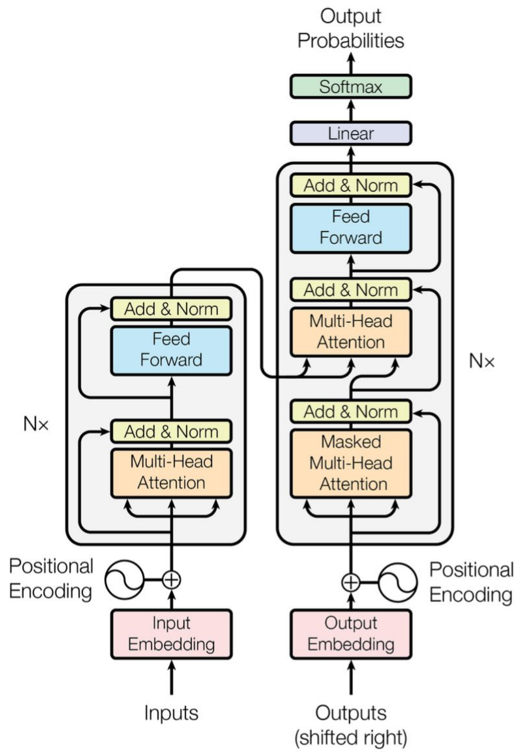


Figure 1: The Transformer - model architecture.

Table Extraction (AWS Textract)

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Label Smoothing During training, we employed label smoothing of value $\epsilon_{ls} = 0.1$ [36]. This hurts perplexity, as the model learns to be more unsure, but improves accuracy and BLEU score.

Data Preparation for Search

- Using the Multimodal LLM (Claude Vision Model), we created descriptions for images and tables. Additionally, we identified figure/table numbers in the images and extracted them in a structured format, creating a dataframe for reference.
- For creating summaries of each thesis, we used Claude(Sonnet 3.5) for its larger context window.
- Once the image/table descriptions summaries has been created, there was no need to use the paid model again for subsequent use of the chatbot.

Figure/Table Description Database Flow

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

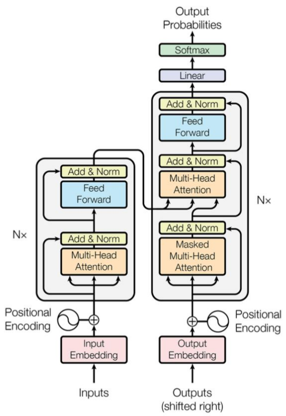
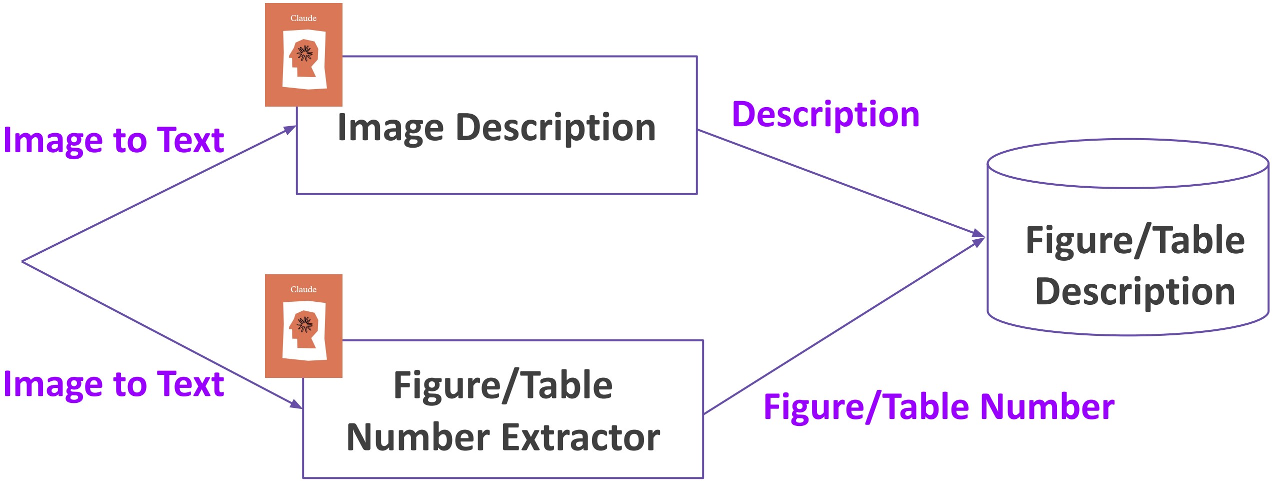


Figure 1: The Transformer - model architecture.



Summary Database Flow

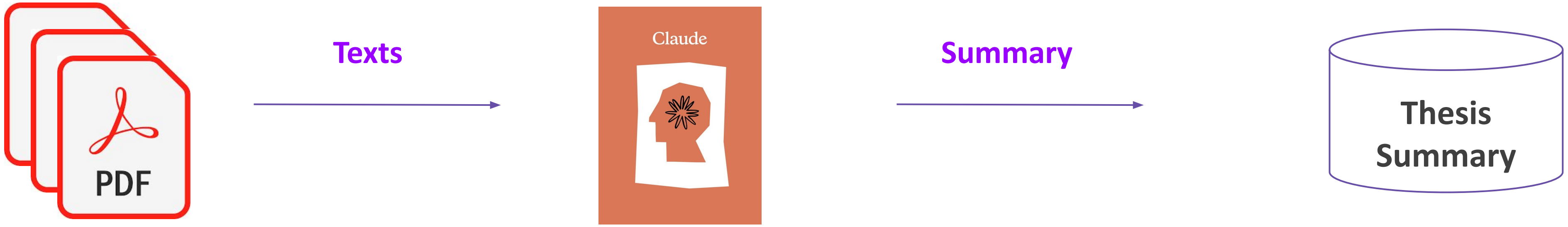


Image Description, Figure/Table Number Extractor

INSIGHT AI

- For image descriptions, we found that the model can provide detailed explanations even for complex images such as Transformer model architecture. By specifying detail instructions in the prompt (e.g., Results and Findings, Visual Aids and Annotations, Limitations and Considerations), we ensured that it could handle a wide range of subsequent questions.
- In the Figure/Table Number Extractor, we refined the few-shot prompting technique. We paid particular attention to restricting the output format to prevent issues with future references.

Image Description

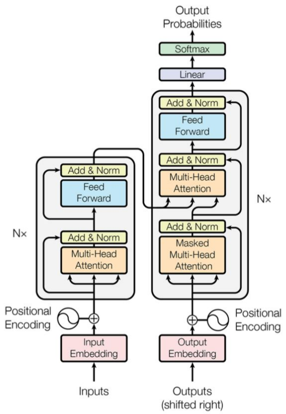


Figure 1: The Transformer - model architecture.

Prompt

Analyze the image… Provide a comprehensive analysis structured as follows…

1. Image Type and Overview:
The image is a flow chart diagram titled…
2. Key Scientific Concepts:
The Transformer is a deep learning model used for natural language processing ….
3. Data Representation:
…the diagram shows the flow of data through the model, with inputs being fed into the encoder…
…
11. Summary
…

Figure/Table Number Extractor

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

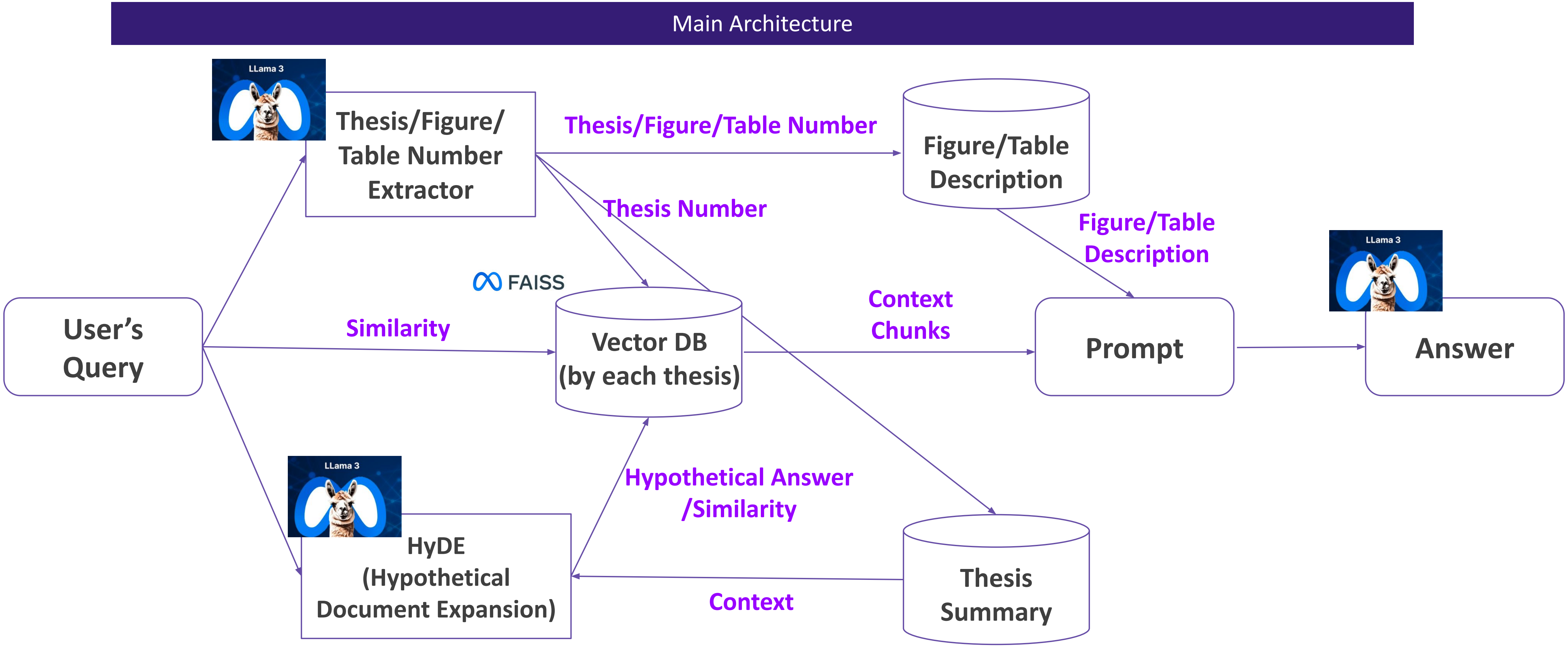
Prompt

You are an engineer skilled in Computer Vision and NLP. Your task is to identify the figure number or table number in the given image. Please respond using the specified format….

"table1"

Main Architecture

- By using a hybrid approach of RAG (General RAG + HyDE) and keyword search (figure/table number), we achieved high performance across a wide range of cases.



Thesis/Figure/Table Number Extractor

- RAG performs searches based on vector similarity, making it well-suited for broad searches and offering high flexibility. However, when specific thesis numbers or figure numbers are provided, the lack of context can result in lower accuracy.
- To address this limitation of RAG, we incorporated exact match searches for these numbers. By using few-shot prompting, we successfully extracted specific numbers from the user's query, ensuring more accurate retrieval of the desired figures or tables.

Prompt

You are an NLP engineer. Your task is to extract the "numbers" from the user's query below.
….
Interpret "figure" for terms such as "Chart," "Diagram," or "Image." Interpret "thesis" for terms such as "Academic Paper," "Paper," or "Document."
….
Please provide your response as a list of objects, each containing thesis, figure, and table.
…

Example1…

Example

Q. Please explain Figure 3 and Table 2 of the second academic paper. What do these indicate about the research findings?

↓

Thesis/Figure/Table Number Extractor

↓

[{"thesis": "2", "figure": "3", "table": ""}, {"thesis": "2", "figure": "", "table": "2" }]

RAG(Retrieval-Augmented Generation)

INSIGHT AI

- In addition to exact match, we use RAG to enable more flexible retrieval that considers the context of the sentences.
- As the embedding model, we use SciBERT, which is specialized for scientific academic papers. Due to the unique vocabulary and expressions often found in academic papers, it is crucial to use an appropriate embedding. By examining the samples below, it is evident that SciBERT retrieves more relevant chunks compared to RoBERTa.
- For the vector database, we use the open-source model FAISS to retrieve the top 5 chunks based on cosine similarity. Additionally, the chunk size is set to 500 characters, with an overlap of 100 to maintain context.

Difference in retrieved chunks using embedding models

Embeddings	Question (for attention.pdf)	Chunk Retrieved
SciBERT (allenai/scibert_scivocab_uncased)	"What is the Attention addressed in the academic article?"	… Figure 5: Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. We give two such examples above, from two different heads from the encoder self-attention at layer 5 of 6. The heads clearly learned to perform different tasks.
RoBERTa (roberta-large)	"What is the Attention addressed in the academic article?"	Content: Computational Linguistics and 44th Annual Meeting of the ACL, pages 433–440. ACL, July 2006. [30] Ofir Press and Lior Wolf. Using the output embedding to improve language models. arXiv preprint arXiv:1608.05859, 2016. …

More relevant

https://github.com/allenai/scibert

RAG - HyDE(Hypothetical Document Expansion)

INSIGHT AI

- HyDE is a method where a hypothetical answer to a question is first generated by an LLM, and then this answer is used to search the VectorDB. In a general RAG setup, we encountered a problem where a user's query was too simple to retrieve highly relevant chunks. To address this, we used HyDE to attempt to retrieve more relevant chunks.
- However, in the context of academic papers, there are many specialized terms that make it difficult to generate even a hypothetical answer without context. To overcome this, we provided the summary of the target thesis as context, allowing for the generation of more meaningful hypothetical answers for the search.
- In the example below, the retrieved chunks with HyDE is actually more relevant to the answers.

Difference in retrieved chunks with and without HyDE				
	Question (for attention.pdf)	Hypothetical Answer	Chunk Retrieved	
With HyDE	"What is the Attention addressed in the academic article?"	The main hypothesis or research question… is whether attention mechanisms alone can replace…..	Attention mechanisms have become an integral part …, allowing modeling of dependencies without regard to their distance in the input or output sequences… however, such attention mechanisms are used in conjunction with a recurrent network. In this work we propose the Transformer, a model architecture eschewing recurrence and instead...	**More relevant**
Without HyDE (General RAG)	"What is the Attention addressed in the academic article?"		… Figure 5: Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. We give two such examples above, from two different heads from the encoder self-attention at layer 5 of 6. The heads clearly learned to perform different tasks.	

Integrated prompt and Final response

- Incorporate descriptions of figures/tables obtained, as well as contexts retrieved by RAG/HyDE, into the prompt.
- Since the chunks retrieved by General RAG and HyDE might differ, both will be used as context. For specialized content like academic papers, it can be difficult to search for information related to the user's query, so we try to maximize recall.
- On the other hand, since accuracy is also crucial for academic papers, the prompt should aim to suppress hallucinations and set the temperature to 0 to minimize diversity.

Prompt

You are an expert in scientific academic papers. Your task is to answer to "Users' query" below.
If the information in the "Figure/Table Context" and "Text Context" below seem relevant to "Users' query", please refer to them.
Please refer only to the relevant contexts for your response.
....

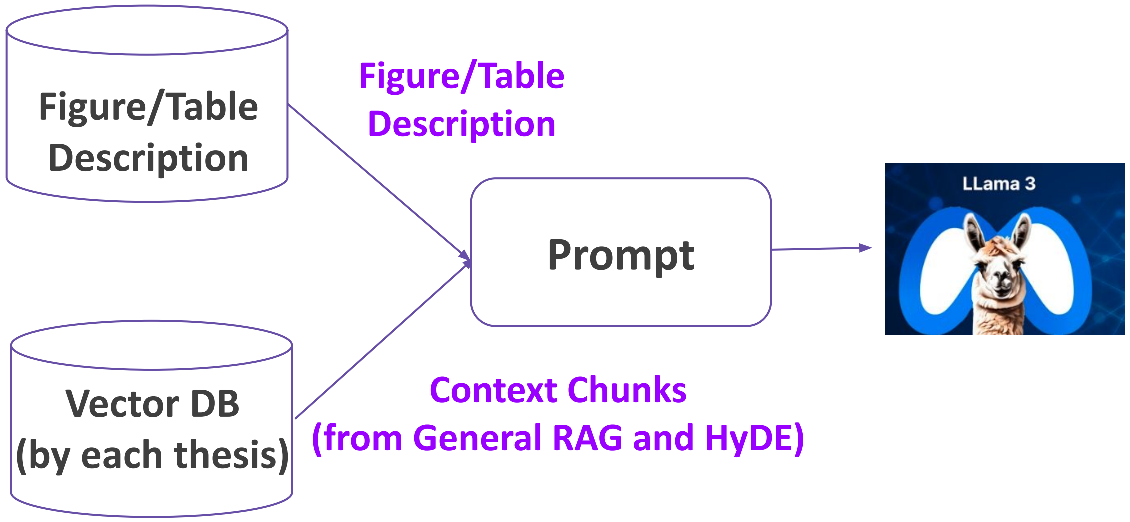
User's query
{USER_QUERY}

Figure/Table Context
{CONTEXT_FIGURE_TABLE}

Text Context
{CONTEXT_RAG_HYDE}

{CONTEXT_RAG_GENERAL}
...

Flow(Simplified)



LIVE DEMO

INSIGHT AI

http://209.97.145.117:8501

Additional Questions for further testing

- In addition to the sample questions provided by the hackathon guidelines, the LLM was able to answer questions about table descriptions and multiple figures from various academic papers while maintaining conversation history. Additionally, it successfully avoided hallucinations.
- For the conversation history(No.2), when conversations span multiple interactions, there is a tendency for the response time to increase and the accuracy to decrease due to the length of the prompts, indicating a need for improvement.

	Questions	Answers	Evaluation
1	Can you explain the table 1 in the first academic paper? **(Table description)** * First paper: attention.pdf	Table 1 in the first paper compares the computational complexity of different neural network layer types commonly used in sequence modeling tasks. The table has 4 columns (Layer Type, Complexity per Layer, Sequential Operations, Maximum Path Length) and 4 rows corresponding to different layer types (Self-Attention, Recurrent, Convolutional, Self-Attention (restricted)). The table provides a valuable reference for comparing the computational efficiency of different neural network layers, highlighting the scalability advantages of self-attention and convolutional layers over recurrent layers.	○
2	Could you give a more concise answer? **(Conversation history)** * It has not been deployed to the production environment.	Here is a more concise answer: Table 1 compares the computational complexity of different neural network layer types (Self-Attention, Recurrent, Convolutional, and restricted Self-Attention) for sequence modeling tasks, highlighting their scalability advantages.	△
3	Describe the Figure1 in the first paper and Figure1 in the second paper. **(Multiple figures from multiple academic papers)** * First paper: attention.pdf * Second paper: Multimodal.pdf	Based on the provided Figure/Table Context, I can describe Figure 1 in the first paper and Figure 1 in the second paper as follows: **Figure 1 in the first paper:** Figure 1 in the first paper illustrates the Transformer model architecture, which consists of an Encoder and a Decoder… **Figure 1 in the second paper:** Figure 1 in the second paper illustrates the LLaVA network architecture, which integrates vision and language…	○
4	Please describe the table 23 first paper. **(Hallucination check)**	The information provided is insufficient to answer your question. Could you please clarify your question? The description of Table 23 is not found in the provided context.	○

Future Work

- **RAG - Improving Chunking Methods**:In this project, we used a fixed-length chunking method, but there is potential for creating more contextually meaningful chunks. Especially in academic papers, subsections or paragraphs often have clear delineations. Adopting such chunking methods could have improved the accuracy of the RAG model.
- **Image Embedding**: In this project, we converted images into text descriptions before embedding. However, this process might result in some information loss, (despite our efforts to minimize this by specifying 11 detailed items for each description.) Exploring direct image embedding could be worthwhile.
- **Fine-tuning**: We used SciBERT, an embedding model specialized for scientific academic papers, but fine-tuning could create a model even more tailored to AI-focused papers. Additionally, while we used a pre-trained LLM, fine-tuning this model could further enhance response accuracy.
- **Adaptation Beyond Typical Academic Papers**: Our primary use case was academic papers, where we retrieved descriptions using figure/table numbers as keys. However, some sample PDFs lacked figure/table numbers. Considering these PDFs also had limited text, integrating text descriptions of each image back into the original PDF before applying RAG could potentially improve accuracy.