

Analysis of Outputs

outputs/outputs.csv allows for easier comparison of the outputs.

outputs/ also contains outputs pertaining solely to each of the models (as referenced below)

LLM (Base Model) -> output_llm.txt

- Provides generic, high-level responses
- Often starts with basic definitions or explanations, giving a broad overview of a topic
- Tends to give theoretical rather than evidence-based answers
- Cannot access data past 2021, and if queried for information past 2022, it gives vague answers

RAG Simple (Incorporation of RAG) -> output_rag_simple.txt

I want to develop a model that is able to give answers that are more academic in general.

- Incorporates more specific examples and data points
- Shows better grounding in recent research
- Still maintains some of the generic structure of LLM responses in that the response is still a little wordy.

RAG Classifier Improvements -> output_rag_classifier.txt

I wanted to ensure that, for more general queries, the RAG would not be too technical.

- More selective about when to use retrieved context
- Better at determining whether academic sources are needed
- Shows improved focus and generally is more concise than the Simple RAG

RAG Rewrite Improvements -> output_rag_rewrite.txt

By rewriting the prompts and then feeding them one by one into the vector DB, I am looking for more of holistic look into the query itself, as often I have found that a question I ask will often lead to followup as the answer provided is either too vague or too focused.

- Consistently starts with "Based on" showing clear grounding in sources
- More sophisticated integration of multiple sources
- Better synthesis of information
- More academic and formal tone in responses

RAG 2 Advanced Improvements

By combining both approaches, I am seeing a consistent increase in quality to the answers that I am getting.

- Combines benefits of both Classifier and Rewrite approaches
- Shows best integration of sources, with a much more holistic response, reducing follow up questions
- Maintains consistent academic tone while providing most detailed and concise responses
- Best at handling complex queries requiring multiple perspectives
- Is able to both handle more general and more detailed questions.