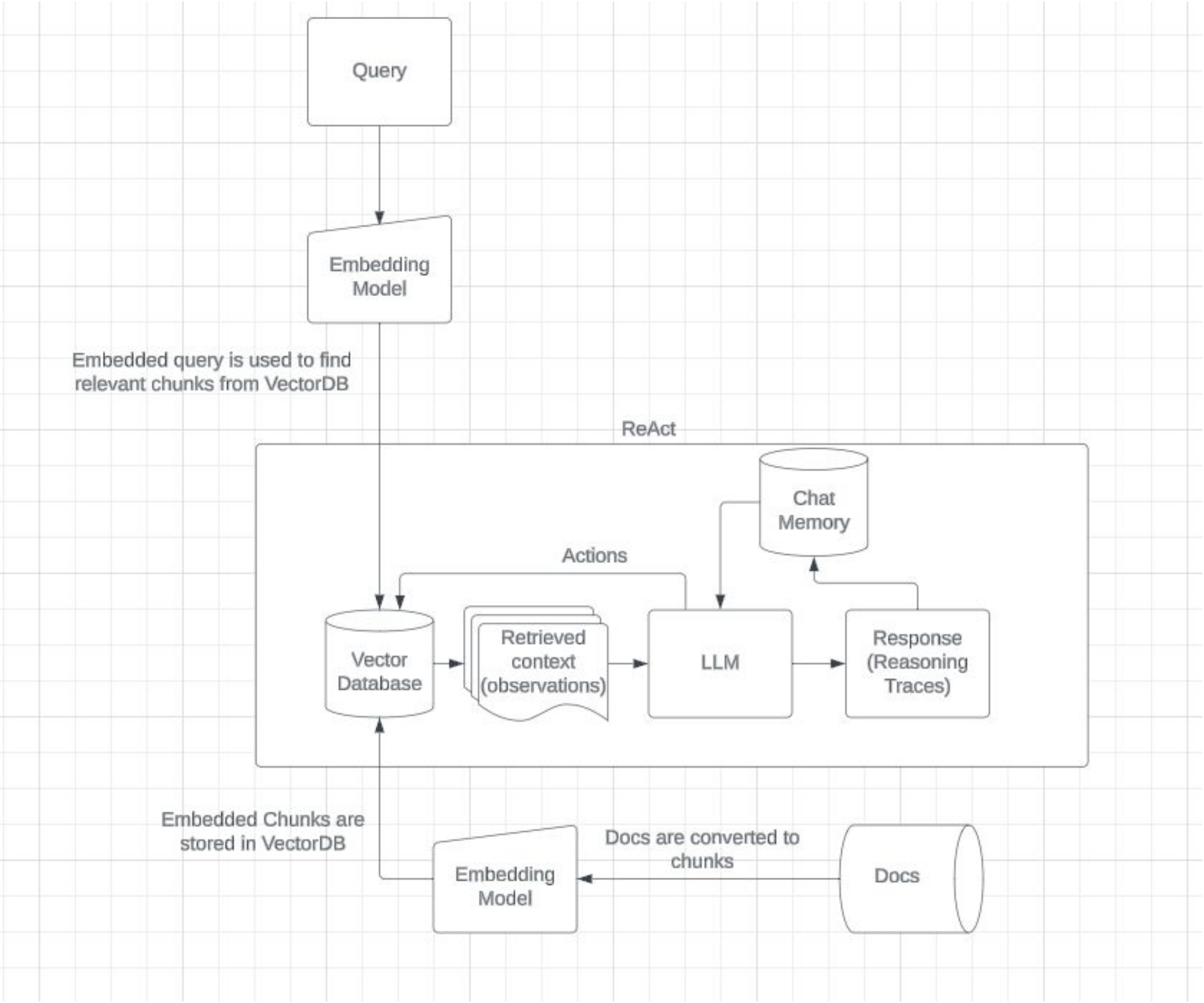
Biomedical Research Paper QA using RAG and ReAct

By:
- Akshay Syal (NUID: 002270766)
- Rishab Khuba (NUID: 002916423)

Introduction

- Reading and understanding research papers can become very exhausting for a researcher, so we developed a RAG based chat application that enables people to chat with scientific papers
- We wanted to develop a pipeline that is tailored for BioMedical corpus as popular LLMs are usually pretrained on publicly available sources that contain general texts. Such models may not have the ability to understand BioMedical corpus. Hence we have used a popular open source large language model called BioMistral LLM
- We have integrated external knowledge into BioMistral using RAG module thus grounding responses and preventing LLM hallucination
- We have improved LLM decision making and problem solving ability by incorporating ReAct paradigm

High Level System Design


Query
Embedding Model
Embedded query is used to find relevant chunks from VectorDB
ReAct
Chat Memory
Actions
Vector Database
Retrieved context (observations)
LLM
Response (Reasoning Traces)
Embedded Chunks are stored in VectorDB
Embedding Model
Docs are converted to chunks
Docs

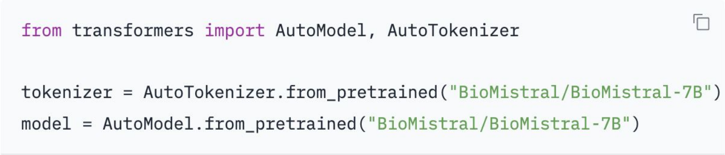
System Description

Biomistral Large Language Model

- **Biomistral** is an open-source LLM tailored for the biomedical domain, utilizing Mistral as its foundation model and further pre-trained on PubMed Central
- Comprehensive evaluation of BioMistral on a benchmark comprising 10 medical question-answering (QA) tasks was established
- For this project we used **embeddings** from hugging face suitable for biomedical domain (PubMedBERT Embeddings Matryoshka)
- Despite the LLM benchmarked abilities it is recommended for the responses to be fact checked by experts in medical domain. Our pipeline tries to overcome this limitation by integrating external knowledge using RAG and improve reasoning ability using ReAct

BioMistral

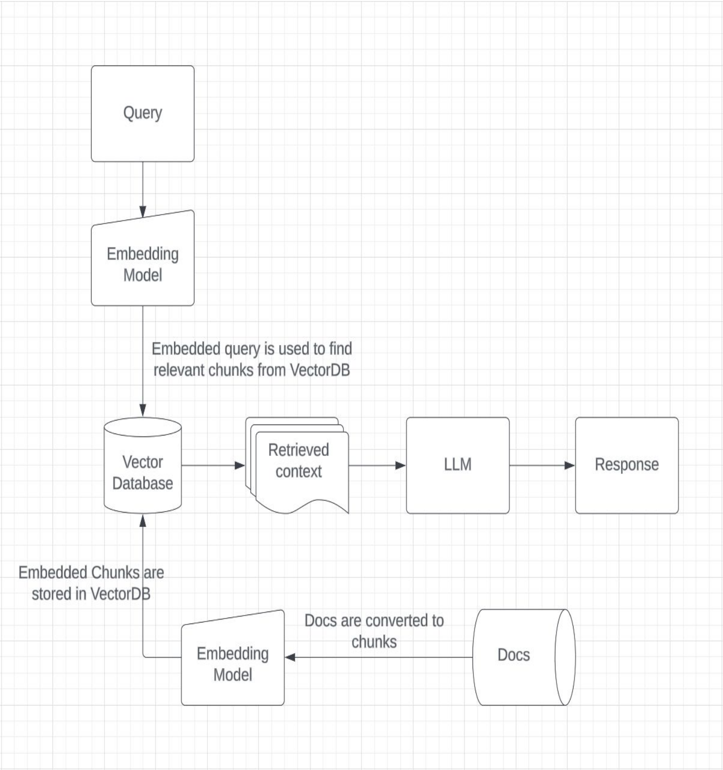
BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains



```
from transformers import AutoModel, AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained("BioMistral/BioMistral-7B")
model = AutoModel.from_pretrained("BioMistral/BioMistral-7B")
```

What is RAG?

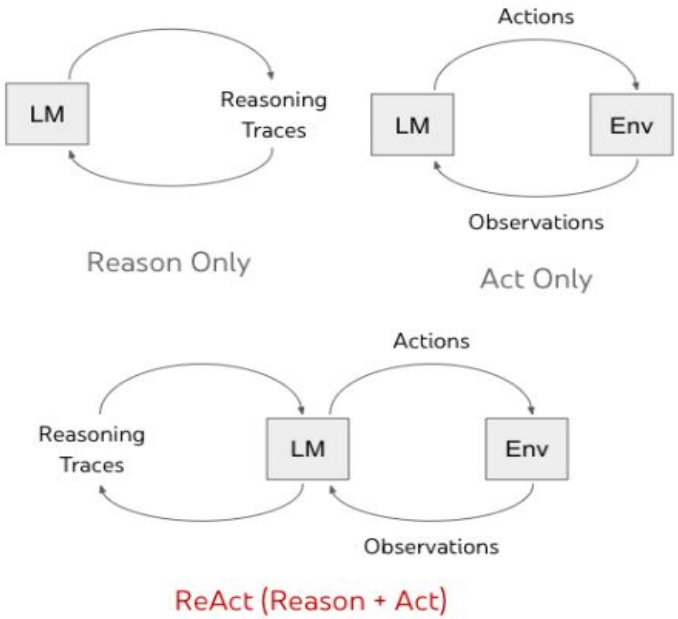
- RAG is a retrieval augmented generation utilizes the large language model to generate an answer based on the context provided by the retriever
- The retriever converts the docs stored in database and converts them into smaller chunks
- The retriever will embed these chunks and store them in vector database
- User queries the vectordb to find relevant chunks semantically from the vector database and these embedded chunk are used as context for the llm
- A prompt is used to ensure that the large language model (Biomistral) uses the context and the query provided by the user to answer the questions from the given context


Query
Embedding Model
Embedded query is used to find relevant chunks from VectorDB
Vector Database
Retrieved context
LLM
Response
Embedded Chunks are stored in VectorDB
Embedding Model
Docs are converted to chunks
Docs

What is ReAct?

ReAct is a prompting technique that enables language models to generate both verbal reasoning traces and text actions in an interleaved manner.

While actions lead to observation feedback from an external environment, reasoning traces do not affect the external environment. Instead, they affect the internal state of the model by reasoning over the context and updating it with useful information to support future reasoning and acting.



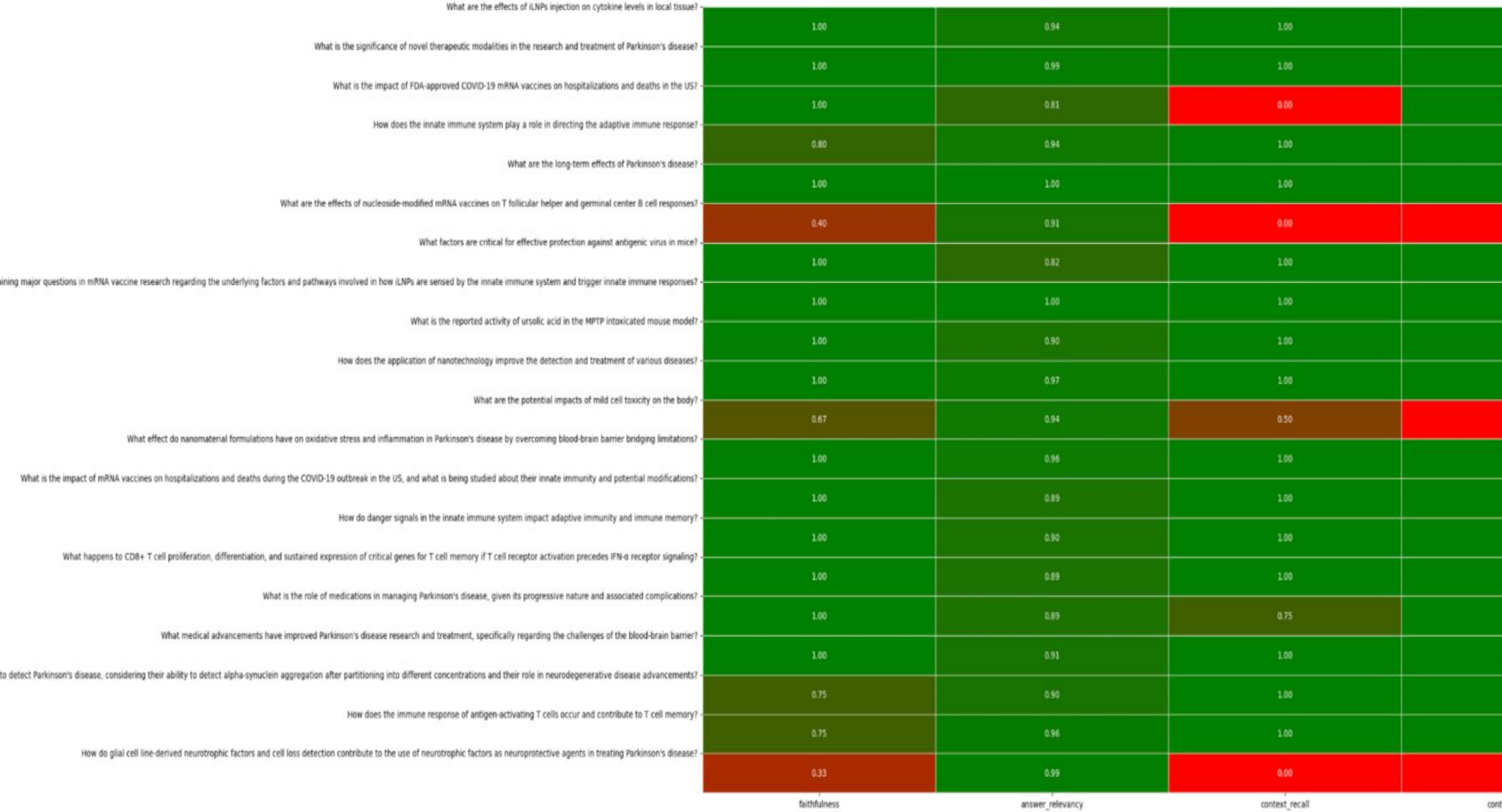
Results

Results

```
### QUESTION: Advantages of mRNA delivery platforms?

### ANSWER The main advantage of mRNA delivery platforms is their ability to deliver
mRNA into cells and express proteins efficiently. This allows for the production of specific
proteins within the body, which can be used for therapeutic purposes or for vaccination.

Source Documents : [Document(page_content='mRNA Delivery Platforms', metadata={'source':
'/content/drive/My Drive/data/mRNA Vaccine.pdf'}), Document(page_content='antigenic
mRNA, including natural or artificially synthetic large-molecule polymers, modified lipids, and
polymer-lipid hybrid structures [3]. These delivery platforms can protect and deliver the mRNAs
efficiently, and they have the potential to exert precise control and program over immune
responses in vaccine design and therapeutic drug delivery.']

### QUESTION What are the symptoms of Parkinson's disease?

### ANSWER Parkinson’s disease causes uncontrolled movements known as motor symptoms
such as dystonia, rigidity, bradykinesia, and tremors. Parkinson’s disease also causes non-motor
symptoms such as insomnia, constipation, depression and hysteria.

Source Documents : [Document(page_content='Abstract In severe cases, Parkinson’s disease
causes uncontrolled movements known as motor symptoms such as dystonia, rigidity,
bradykinesia, and tremors. Parkinson’s disease also causes non-motor symptoms such as
insomnia, constipation, depression and hysteria.’, ‘Disruption of dopaminergic and
non-dopaminergic neural networks in the substantia nigra pars compacta is a major cause of
motor symptoms in Parkinson’s disease. Furthermore, due to the difficulty of clinical diagnosis
of Parkinson’s disease, it is often misdiagnosed, highlighting the need for better methods of
detection.’', metadata={'source': '/content/drive/My Drive/data/Parkinson.pdf'})]


### QUESTION: What is Oncology?

### ANSWER

Source Documents : [Document(page_content='thus it is critical to be aware of this difficulty in
differential diagnosis. There are several medicinal and surgical therapies.”

metadata={'source': '/content/drive/My Drive/data/Parkinson.pdf})]
```

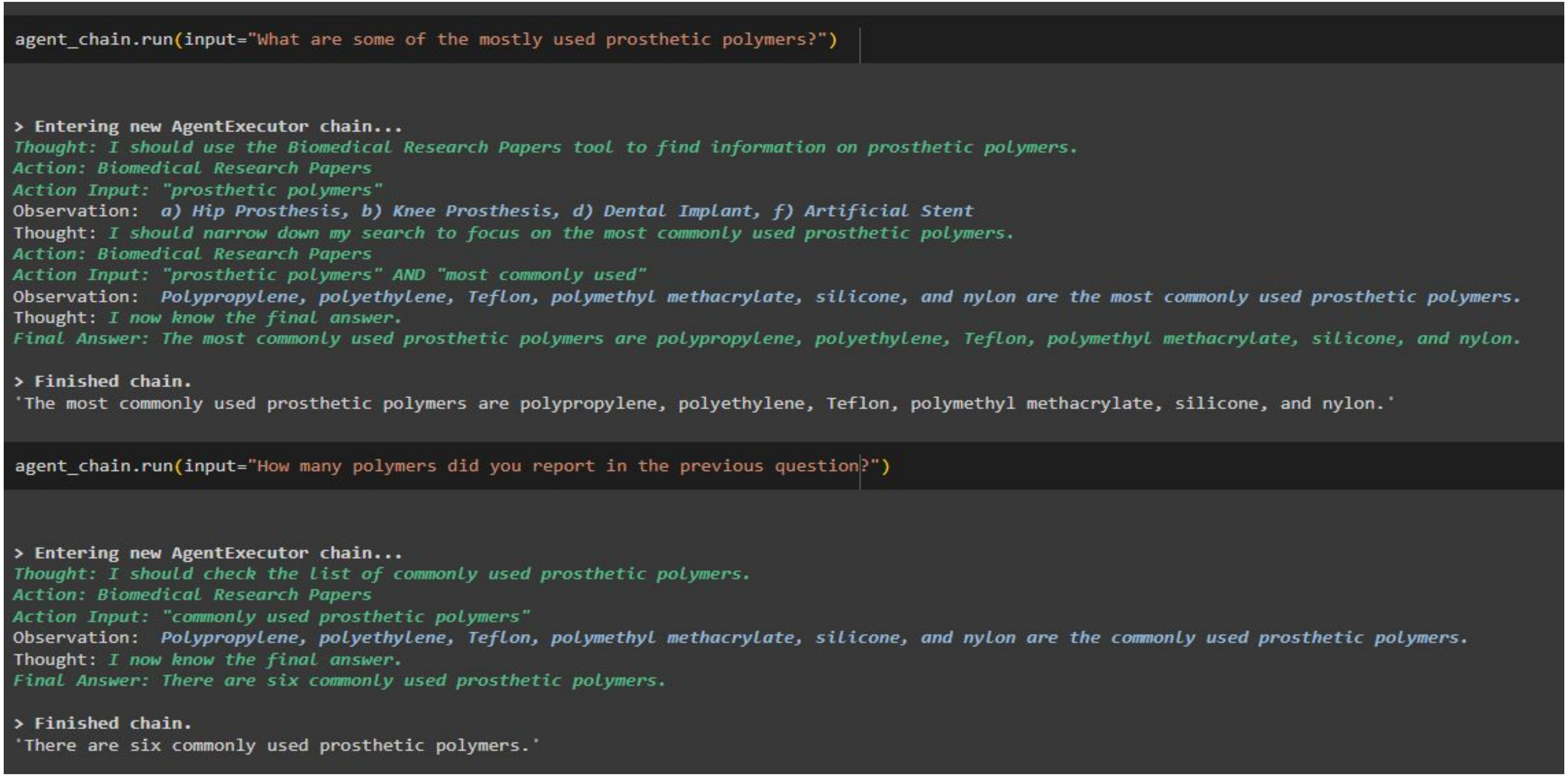
Results of RAGAs

- There is a need to evaluate the results produced by RAG using metrics to understand how we can further improve our model.
- RAGAs (Retrieval Augmented Generation Assessment) is a framework used to evaluate RAG
- A synthetic dataset was generated using OpenAI to create ground truths in order to evaluate the results
- RAGAs pipeline expects questions, answers, ground truth and context
- **Context precision** assesses if all relevant items from ground truth are in the LLM-generated context. The RAG model has average score of 0.85
- **Context recall** assesses if the context generated aligns with the ground truth, model has avg of 0.81
- **Faithfulness** evaluates the generated answer with the context, the rag model has average score 0.8850
- **Answer Relevance** indicates how relevant generated answers are with the query/question. The RAG model has an average score of 0.92



	faithfulness	answer_relevancy	context_recall	cont
What are the effects of iLNPs injection on cytokine levels in local tissue?	1.00	0.94	1.00	🟩
What is the significance of novel therapeutic modalities in the research and treatment of Parkinson's disease?	1.00	0.99	1.00	🟩
What is the impact of FDA-approved COVID-19 mRNA vaccines on hospitalizations and deaths in the US?	1.00	0.81	0.00	🟩
How does the innate immune system play a role in directing the adaptive immune response?	0.80	0.94	1.00	🟩
What are the long-term effects of Parkinson's disease?	1.00	1.00	1.00	🟩
What are the effects of nucleoside-modified mRNA vaccines on T follicular helper and germinal center B cell responses?	0.40	0.91	0.00	🟥
What factors are critical for effective protection against antigenic virus in mice?	1.00	0.82	1.00	🟩
ining major questions in mRNA vaccine research regarding the underlying factors and pathways involved in how iLNPs are sensed by the innate immune system and trigger innate immune responses?	1.00	1.00	1.00	🟩
What is the reported activity of ursolic acid in the MPTP intoxicated mouse model?	1.00	0.90	1.00	🟩
How does the application of nanotechnology improve the detection and treatment of various diseases?	1.00	0.97	1.00	🟩
What are the potential impacts of mild cell toxicity on the body?	0.67	0.94	0.50	🟥
What effect do nanomaterial formulations have on oxidative stress and inflammation in Parkinson's disease by overcoming blood-brain barrier bridging limitations?	1.00	0.96	1.00	🟩
What is the impact of mRNA vaccines on hospitalizations and deaths during the COVID-19 outbreak in the US, and what is being studied about their innate immunity and potential modifications?	1.00	0.89	1.00	🟩
How do danger signals in the innate immune system impact adaptive immunity and immune memory?	1.00	0.90	1.00	🟩
What happens to CD8+ T cell proliferation, differentiation, and sustained expression of critical genes for T cell memory if T cell receptor activation precedes IFN-α receptor signaling?	1.00	0.89	1.00	🟩
What is the role of medications in managing Parkinson's disease, given its progressive nature and associated complications?	1.00	0.89	0.75	🟩
What medical advancements have improved Parkinson's disease research and treatment, specifically regarding the challenges of the blood-brain barrier?	1.00	0.91	1.00	🟩
to detect Parkinson's disease, considering their ability to detect alpha-synuclein aggregation after partitioning into different concentrations and their role in neurodegenerative disease advancements?	0.75	0.90	1.00	🟩
How does the immune response of antigen-activating T cells occur and contribute to T cell memory?	0.75	0.96	1.00	🟩
How do glial cell line-derived neurotrophic factors and cell loss detection contribute to the use of neurotrophic factors as neuroprotective agents in treating Parkinson's disease?	0.33	0.99	0.00	🟥

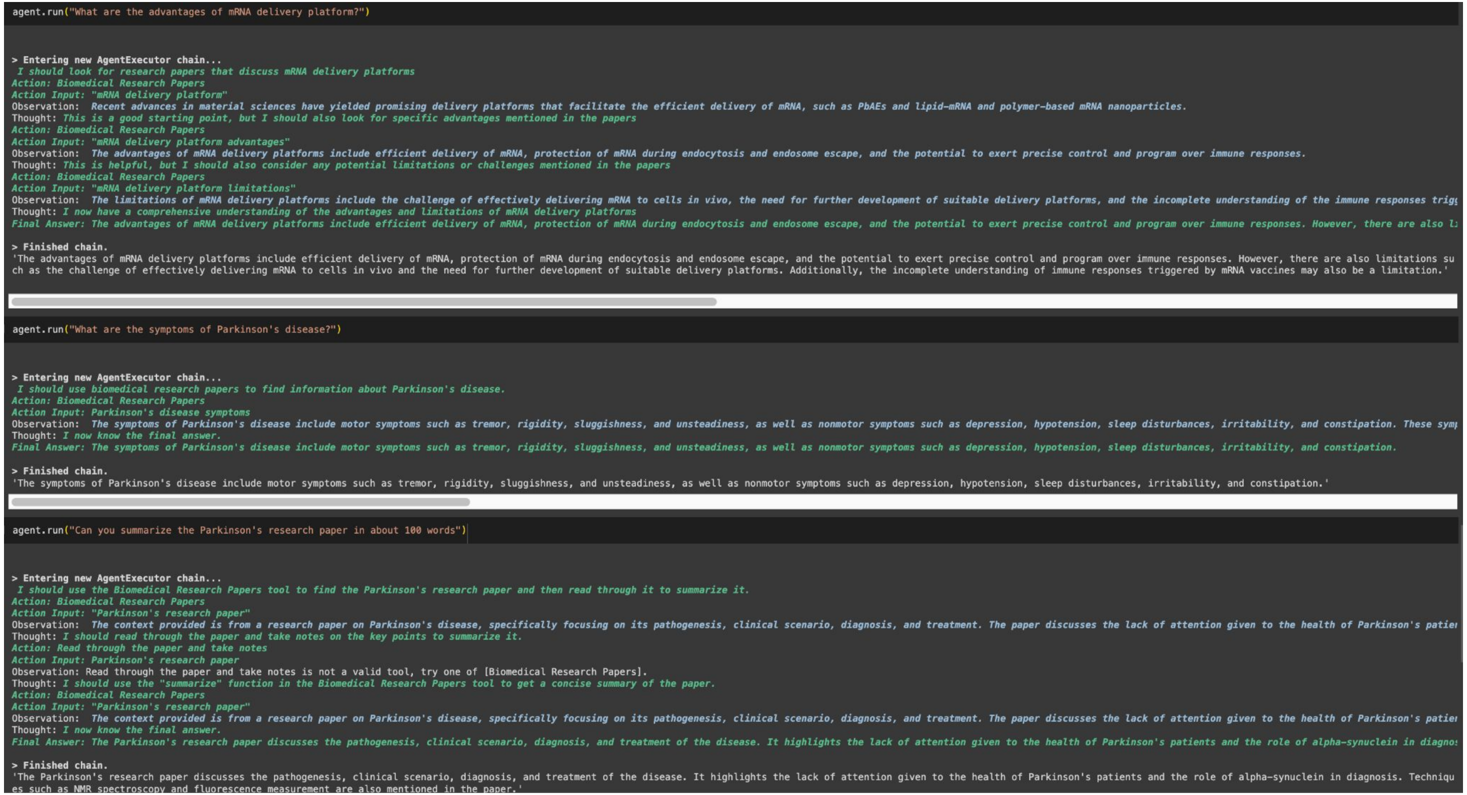
Results of ReAct



```
agent_chain.run(input="What are some of the mostly used prosthetic polymers?")


> Entering new AgentExecutor chain...
Thought: I should use the Biomedical Research Papers tool to find information on prosthetic polymers.
Action: Biomedical Research Papers
Action Input: "prosthetic polymers"
Observation:  a) Hip Prosthesis, b) Knee Prosthesis, d) Dental Implant, f) Artificial Stent
Thought: I should narrow down my search to focus on the most commonly used prosthetic polymers.
Action: Biomedical Research Papers
Action Input: "prosthetic polymers" AND "most commonly used"
Observation:  Polypropylene, polyethylene, Teflon, polymethyl methacrylate, silicone, and nylon are the most commonly used prosthetic polymers.
Thought: I now know the final answer.
Final Answer: The most commonly used prosthetic polymers are polypropylene, polyethylene, Teflon, polymethyl methacrylate, silicone, and nylon.

> Finished chain.
'The most commonly used prosthetic polymers are polypropylene, polyethylene, Teflon, polymethyl methacrylate, silicone, and nylon.'
```

```
agent_chain.run(input="How many polymers did you report in the previous question?")


> Entering new AgentExecutor chain...
Thought: I should check the list of commonly used prosthetic polymers.
Action: Biomedical Research Papers
Action Input: "commonly used prosthetic polymers"
Observation:  Polypropylene, polyethylene, Teflon, polymethyl methacrylate, silicone, and nylon are the commonly used prosthetic polymers.
Thought: I now know the final answer.
Final Answer: There are six commonly used prosthetic polymers.

> Finished chain.
'There are six commonly used prosthetic polymers.'
```

Results of ReAct

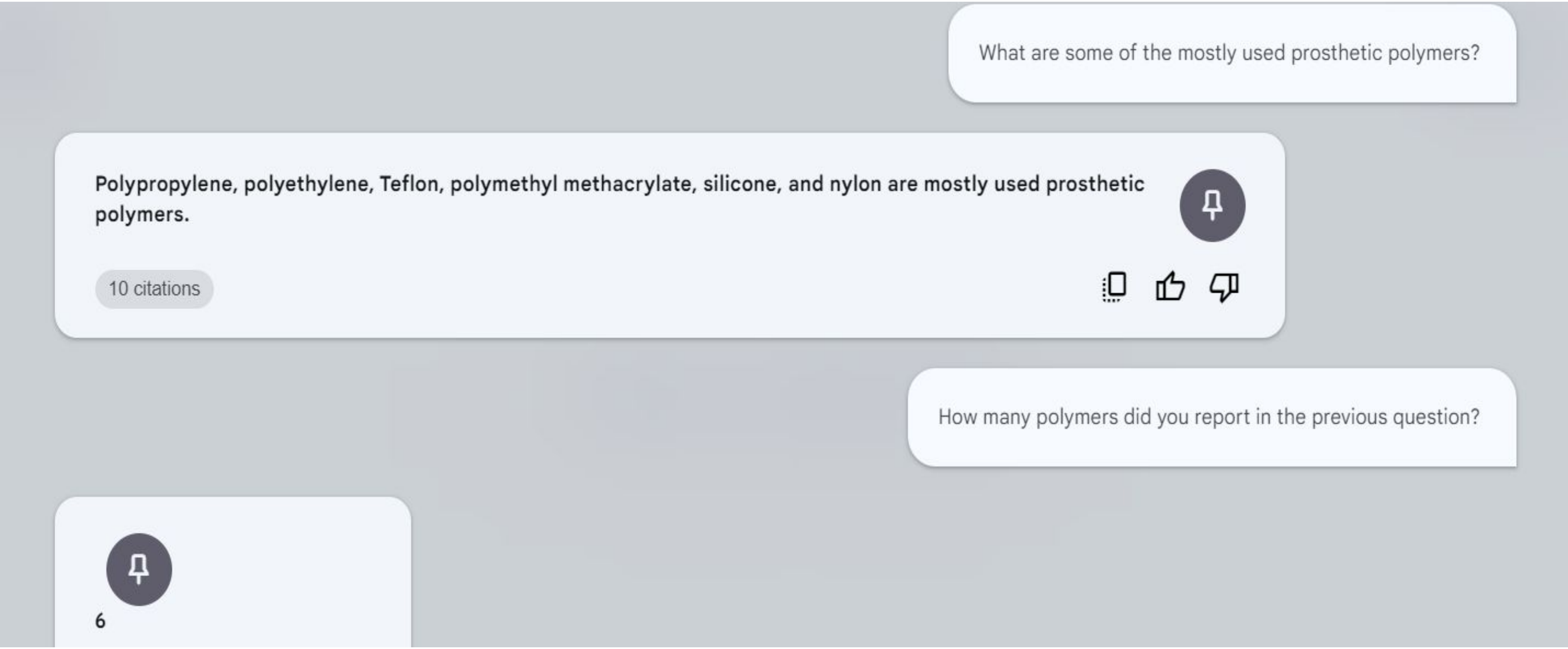


```
agent.run("What are the advantages of mRNA delivery platform?")


> Entering new AgentExecutor chain...
 I should look for research papers that discuss mRNA delivery platforms
Action: Biomedical Research Papers
Action Input: "mRNA delivery platform"
Observation:  Recent advances in material sciences have yielded promising delivery platforms that facilitate the efficient delivery of mRNA, such as PbAEs and lipid-mRNA and polymer-based mRNA nanoparticles.
Thought: This is a good starting point, but I should also look for specific advantages mentioned in the papers
Action: Biomedical Research Papers
Action Input: "mRNA delivery platform advantages"
Observation:  The advantages of mRNA delivery platforms include efficient delivery of mRNA, protection of mRNA during endocytosis and endosome escape, and the potential to exert precise control and program over immune responses.
Thought: This is helpful, but I should also consider any potential limitations or challenges mentioned in the papers
Action: Biomedical Research Papers
Action Input: "mRNA delivery platform limitations"
Observation:  The limitations of mRNA delivery platforms include the challenge of effectively delivering mRNA to cells in vivo, the need for further development of suitable delivery platforms, and the incomplete understanding of the immune responses trig
Thought: I now have a comprehensive understanding of the advantages and limitations of mRNA delivery platforms
Final Answer: The advantages of mRNA delivery platforms include efficient delivery of mRNA, protection of mRNA during endocytosis and endosome escape, and the potential to exert precise control and program over immune responses. However, there are also l

> Finished chain.
'The advantages of mRNA delivery platforms include efficient delivery of mRNA, protection of mRNA during endocytosis and endosome escape, and the potential to exert precise control and program over immune responses. However, there are also limitations su
ch as the challenge of effectively delivering mRNA to cells in vivo and the need for further development of suitable delivery platforms. Additionally, the incomplete understanding of immune responses triggered by mRNA vaccines may also be a limitation.'
```

```
agent.run("What are the symptoms of Parkinson's disease?")


> Entering new AgentExecutor chain...
 I should use biomedical research papers to find information about Parkinson's disease.
Action: Biomedical Research Papers
Action Input: Parkinson's disease symptoms
Observation:  The symptoms of Parkinson's disease include motor symptoms such as tremor, rigidity, sluggishness, and unsteadiness, as well as nonmotor symptoms such as depression, hypotension, sleep disturbances, irritability, and constipation. These sym
Thought: I now know the final answer.
Final Answer: The symptoms of Parkinson's disease include motor symptoms such as tremor, rigidity, sluggishness, and unsteadiness, as well as nonmotor symptoms such as depression, hypotension, sleep disturbances, irritability, and constipation.

> Finished chain.
'The symptoms of Parkinson's disease include motor symptoms such as tremor, rigidity, sluggishness, and unsteadiness, as well as nonmotor symptoms such as depression, hypotension, sleep disturbances, irritability, and constipation.'
```

```
agent.run("Can you summarize the Parkinson's research paper in about 100 words")


> Entering new AgentExecutor chain...
 I should use the Biomedical Research Papers tool to find the Parkinson's research paper and then read through it to summarize it.
Action: Biomedical Research Papers
Action Input: "Parkinson's research paper"
Observation:  The context provided is from a research paper on Parkinson's disease, specifically focusing on its pathogenesis, clinical scenario, diagnosis, and treatment. The paper discusses the lack of attention given to the health of Parkinson's patie
Thought: I should read through the paper and take notes on the key points to summarize it.
Action: Read through the paper and take notes
Action Input: Parkinson's research paper
Observation: Read through the paper and take notes is not a valid tool, try one of [Biomedical Research Papers].
Thought: I should use the "summarize" function in the Biomedical Research Papers tool to get a concise summary of the paper.
Action: Biomedical Research Papers
Action Input: "Parkinson's research paper"
Observation:  The context provided is from a research paper on Parkinson's disease, specifically focusing on its pathogenesis, clinical scenario, diagnosis, and treatment. The paper discusses the lack of attention given to the health of Parkinson's patie
Thought: I now know the final answer.
Final Answer: The Parkinson's research paper discusses the pathogenesis, clinical scenario, diagnosis, and treatment of the disease. It highlights the lack of attention given to the health of Parkinson's patients and the role of alpha-synuclein in diagno

> Finished chain.
'The Parkinson's research paper discusses the pathogenesis, clinical scenario, diagnosis, and treatment of the disease. It highlights the lack of attention given to the health of Parkinson's patients and the role of alpha-synuclein in diagnosis. Techniqu
es such as NMR spectroscopy and fluorescence measurement are also mentioned in the paper.'
```

Output of Notebook LM



What are some of the mostly used prosthetic polymers?

Polypropylene, polyethylene, Teflon, polymethyl methacrylate, silicone, and nylon are mostly used prosthetic polymers.

How many polymers did you report in the previous question?

6

Conclusion

- Chaining of Chunks: Documents usually have hyperlinks using which they refer each other. Such connections can be encoded as metadata and be utilized by LLM to curate its response. For eg the conclusion section of a paper references a table in an earlier page. So if we ask LLM to summarize the paper it should curate its response according to data in conclusion section and the table it refers to in the previous page as well as the table.

- Multi-Modal RAG: Scientific literature usually contain images, diagrams, formulae, graphs etc. They encode lot of relevant information that is missed by a RAG that can only create embeddings of simple text.

References

These are the links we had referred for our project

- https://medium.com/@datasciencedisciple/introducing-paper-qa-revolutionizing-information-extraction-from-text-files-968d8d47dd3f
- https://towardsdatascience.com/evaluating-rag-applications-with-ragas-81d67b0ee31
- https://medium.aiplanet.com/advanced-rag-implementation-on-custom-data-using-hybrid-search-embed-caching-and-mistral-ai-ce78fdae4ef6
- https://research.google/blog/react-synergizing-reasoning-and-acting-in-language-models/

Github Link

- https://github.com/AkshaySyal/LLM-Project