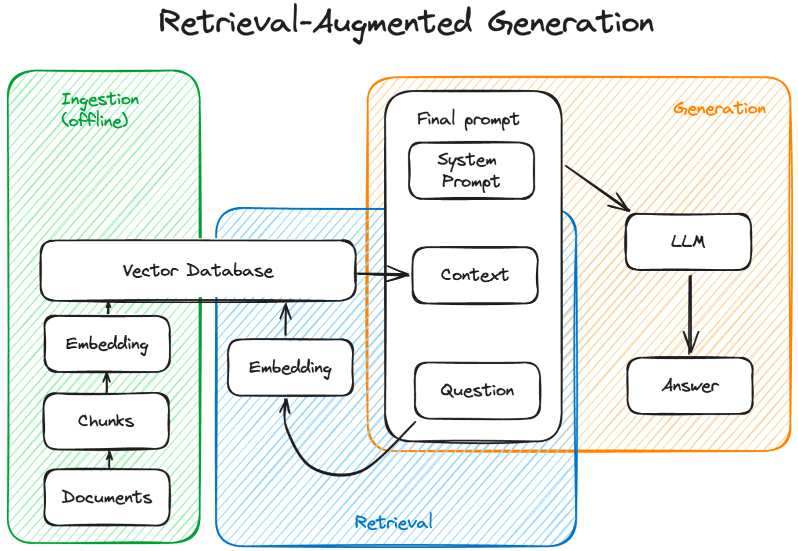
LLM Intelligent Document QA System

By: Nazila Gundogdu

October 2025

Project Goal

Use **Retrieval-Augmented Generation (RAG)**-based LLM to answer questions based on long, unstructured documents, by retrieving relevant chunks.



Data: Alzheimer's Research

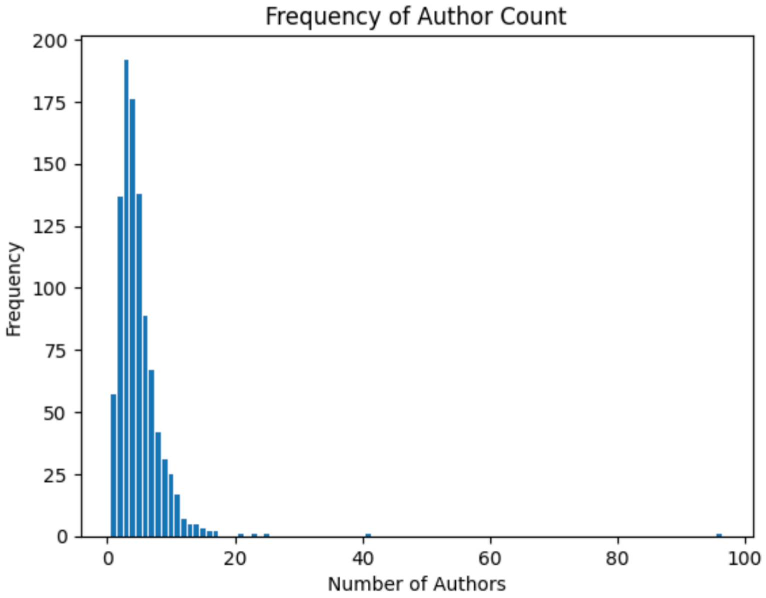
- Focus on **Alzheimer's disease** papers from arXiv.org
- Limited to **1,000 entries**
- Only **abstracts** used
- Collected via invoking arXiv.org API by using `parse.py` script

Data Wrangling and EDA

- Tokenized text and removed stop words using `nltk`
- Analyzed word frequency, author count, token length and publications per year

Sample Insight

- Average authors per paper: **~5**
- Some papers had up to **96 authors!**



Preprocessing & Modeling

LLM Used:

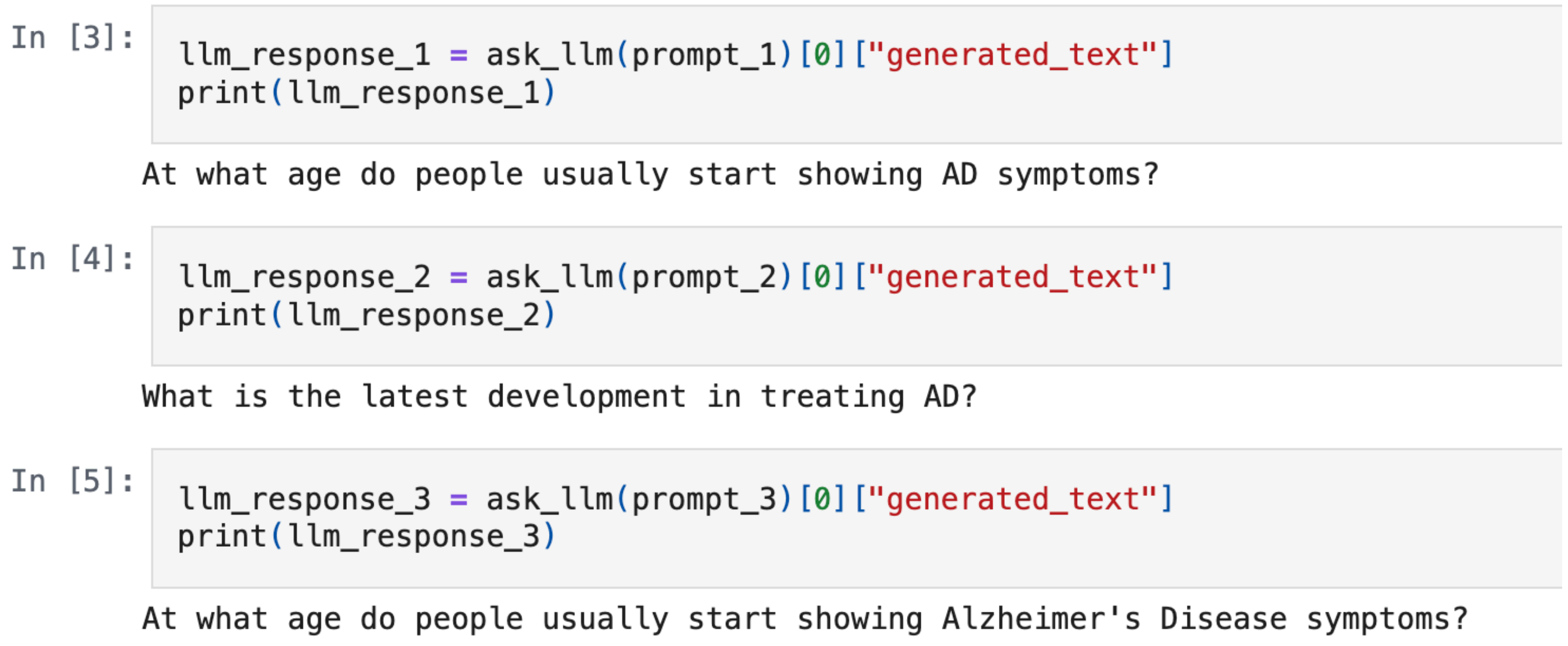
- `TinyLlama-1.1B-Chat-v1.0`

Sample Prompts:

```
prompt_1 = "At what age do people usually start showing AD symptoms?"
prompt_2 = "What is the latest development in treating AD?"
prompt_3 = "At what age do people usually start showing Alzheimer's Disease symptoms?"
prompt_4 = prompt_3 + " Give me a number."
```

Evaluation: RAG vs Non-RAG

Answers Without RAG



```
In [3]:
llm_response_1 = ask_llm(prompt_1)[0]["generated_text"]
print(llm_response_1)
```

```
At what age do people usually start showing AD symptoms?
```

```
In [4]:
llm_response_2 = ask_llm(prompt_2)[0]["generated_text"]
print(llm_response_2)
```

```
What is the latest development in treating AD?
```

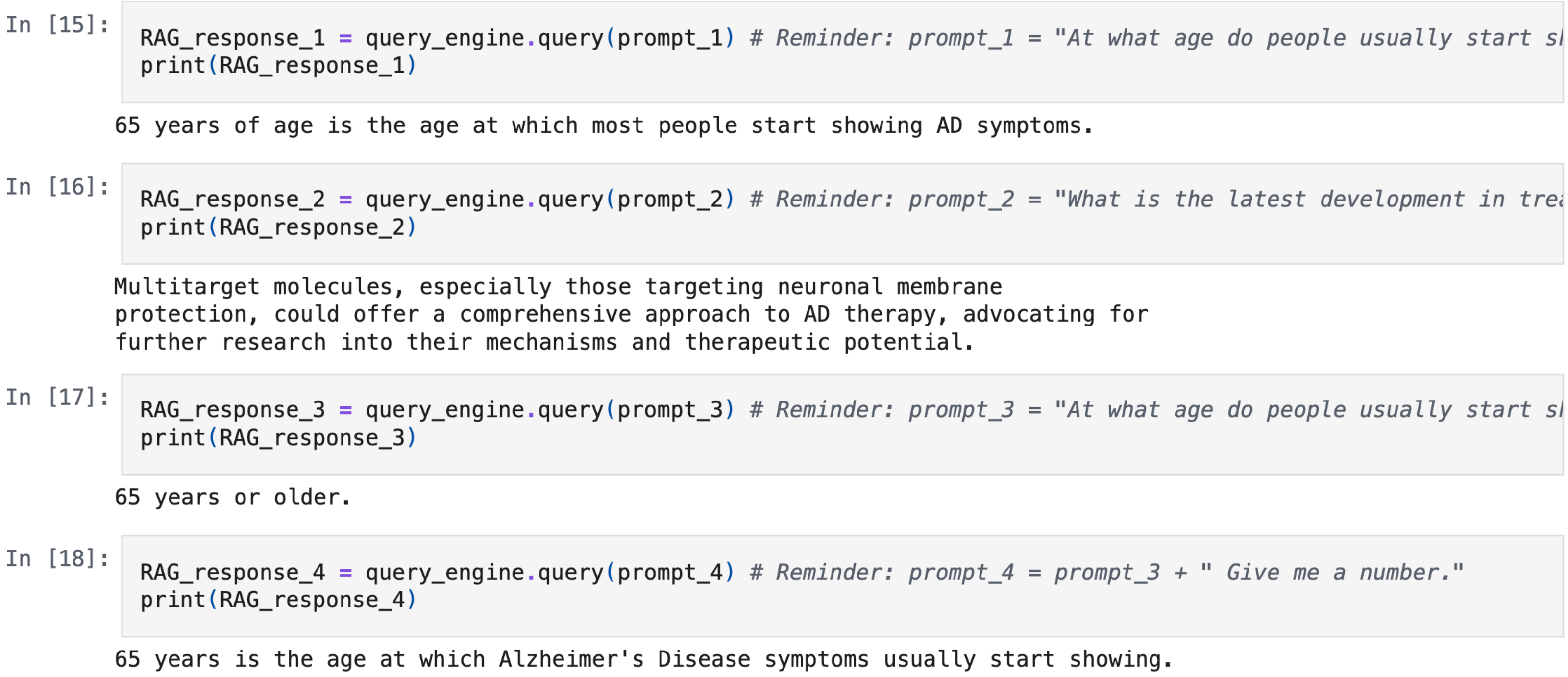
```
In [5]:
llm_response_3 = ask_llm(prompt_3)[0]["generated_text"]
print(llm_response_3)
```

```
At what age do people usually start showing Alzheimer's Disease symptoms?
```

Preprocessing Pipeline

- Convert abstracts to `Document` format
- Split using `SentenceSplitter` (LlamaIndex)
- Vectorize using `all-MiniLM-L6-v2` (Hugging Face)

Answers With RAG



```
In [15]: RAG_response_1 = query_engine.query(prompt_1) # Reminder: prompt_1 = "At what age do people usually start sl
         print(RAG_response_1)
```

```
65 years of age is the age at which most people start showing AD symptoms.
```

```
In [16]: RAG_response_2 = query_engine.query(prompt_2) # Reminder: prompt_2 = "What is the latest development in trea
         print(RAG_response_2)
```

```
Multitarget molecules, especially those targeting neuronal membrane
protection, could offer a comprehensive approach to AD therapy, advocating for
further research into their mechanisms and therapeutic potential.
```

```
In [17]: RAG_response_3 = query_engine.query(prompt_3) # Reminder: prompt_3 = "At what age do people usually start sl
         print(RAG_response_3)
```

```
65 years or older.
```

```
In [18]: RAG_response_4 = query_engine.query(prompt_4) # Reminder: prompt_4 = prompt_3 + " Give me a number."
         print(RAG_response_4)
```

```
65 years is the age at which Alzheimer's Disease symptoms usually start showing.
```

Limitations

- Used small models and limited data
- No prompt engineering

Results & Future Work

- RAG improves accuracy and grounding
- Scale to larger models and datasets
- Experiment with:
  - Prompt engineering
  - Context formatting
  - Evaluation/comparison methods
- Deploy in real-world QA settings