Retrieval Augmented Generation

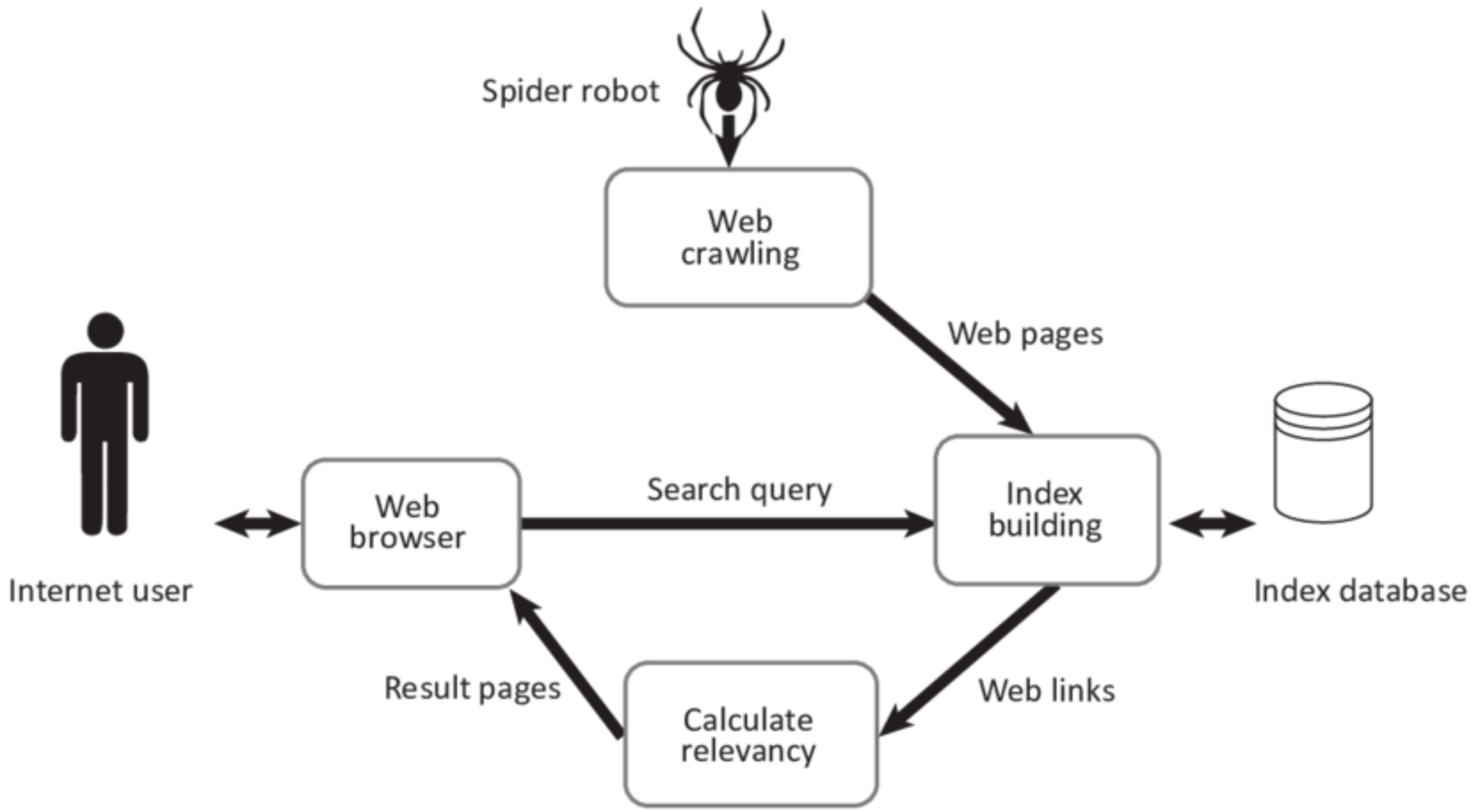
Merging LLMs with IR

Parsa Toopchinezhad - Fall 2023 - Information Retrieval - Dr. Monkaresi

Search Engines

1990s - present

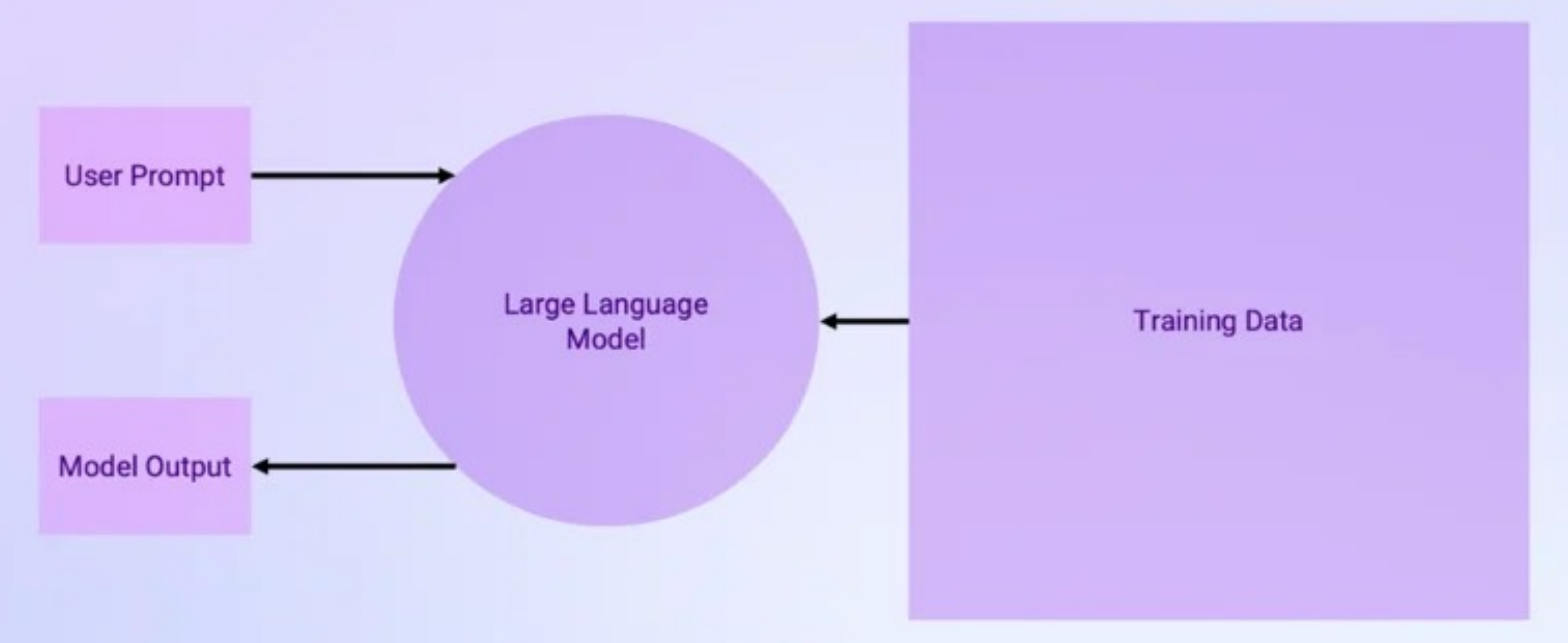
- The main method of search since the 90s
- The topic of this course
- IR find results through millions of webpages



Generative Language Models

2022 - present

- Became mainstream in 2022 thanks to ChatGPT
- Massive ML models trained on huge datasets
- Generate instead of searching



Traditional IR v.s LLMs

Advantages

- Wide variety of applications
- Human like response
- Can understand basic logic
- Ease of use (no more hauling through websites)

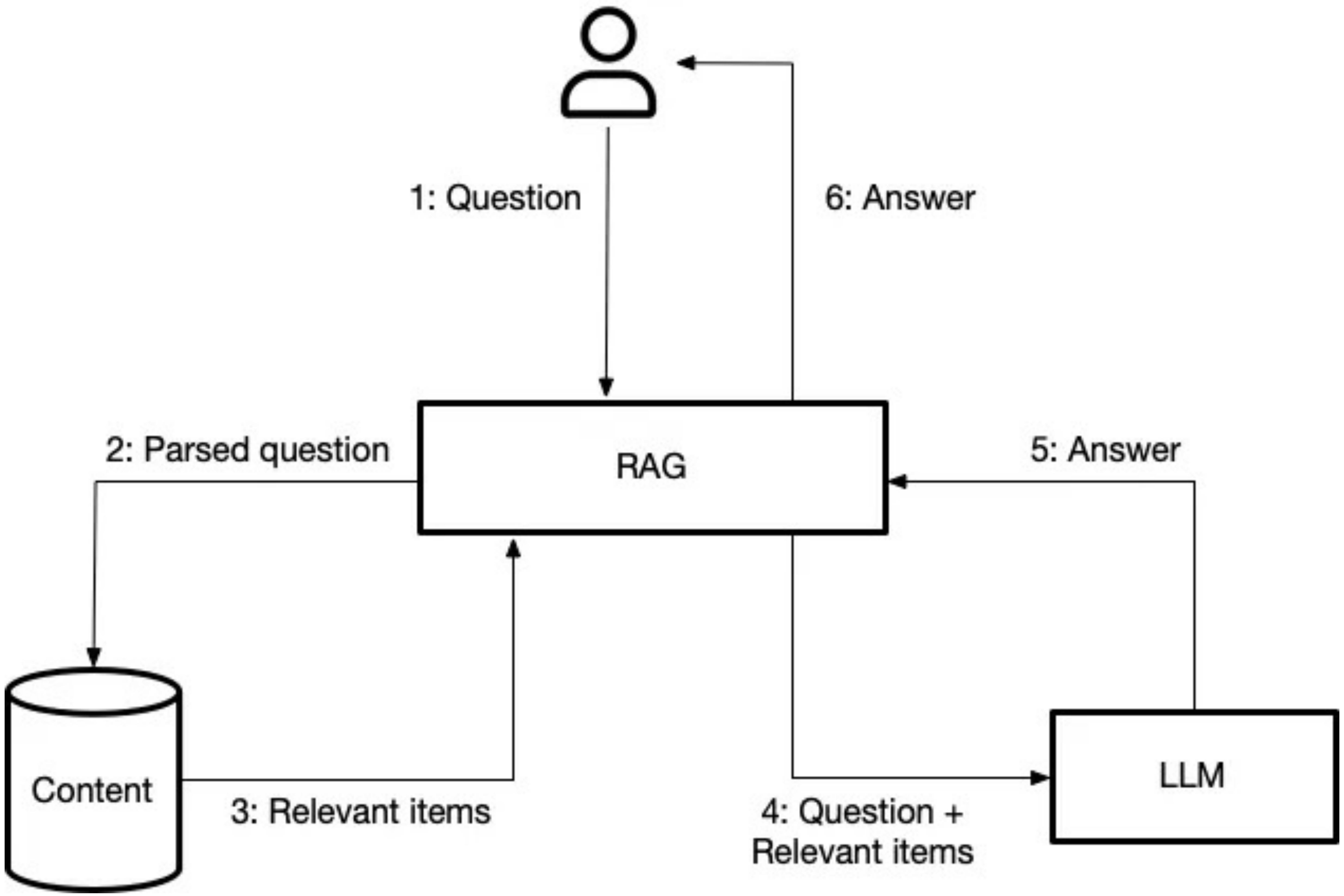
Disadvantages

- Hallucinations
- No source for answers
- Become outdated with time

What if we could combine the two?

Retrieval Augmented Generation (RAG)

IR meets LLMs



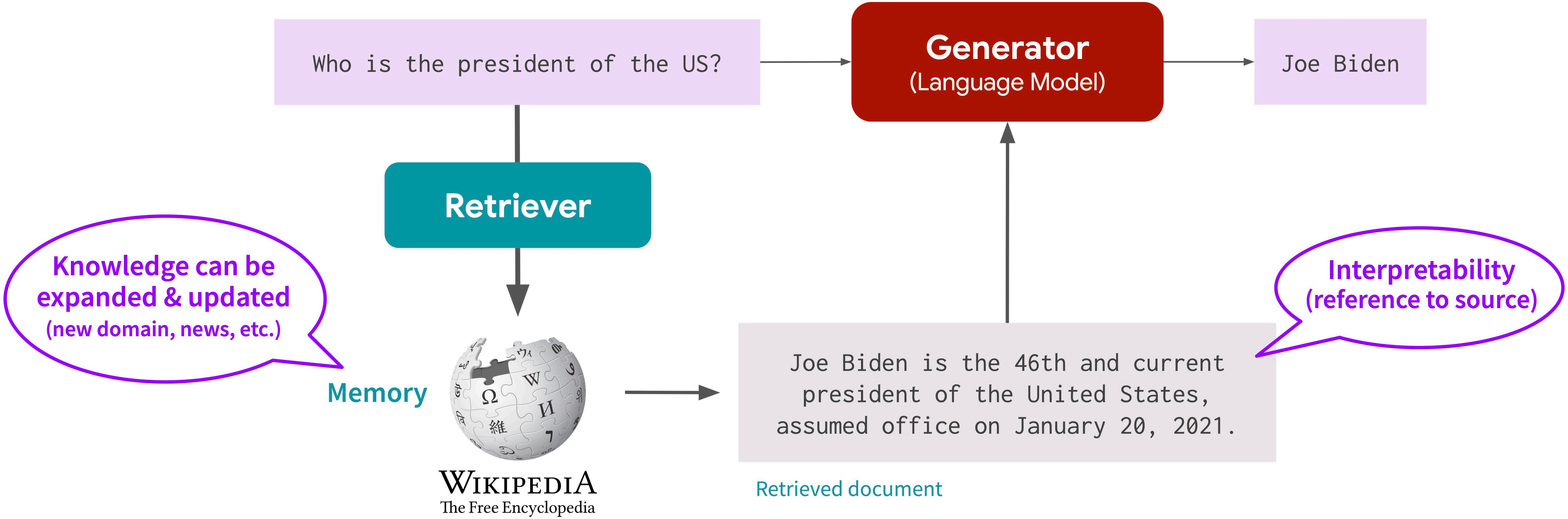
RAG

A formal introduction

- A relatively new method for Grounding LLMs on external sources of knowledge
- Greatly enhances the accuracy and reliability of generative AI models
- Combines the strengths of generative and retrieval models
- Lower rates of hallucinations

RAG Example

Retrieval augmentation



Who is the president of the US?

Generator
(Language Model)

Joe Biden

Retriever

Knowledge can be expanded & updated
(new domain, news, etc.)

Memory

WIKIPEDIA
The Free Encyclopedia

Joe Biden is the 46th and current president of the United States, assumed office on January 20, 2021.

Retrieved document

Interpretability
(reference to source)

RAG Applications

- Getting up to date information
  - Real time statistics/news
- Using LLM for private data
  - Company data
  - Personal data
- Creating specialized LLMs
  - Healthcare
  - Education
  - E-commerce

References

- https://jettro.dev/question-answering-through-retrieval-augmented-generation-9b54806c214e