Retrieval Augmented Generation (RAG)

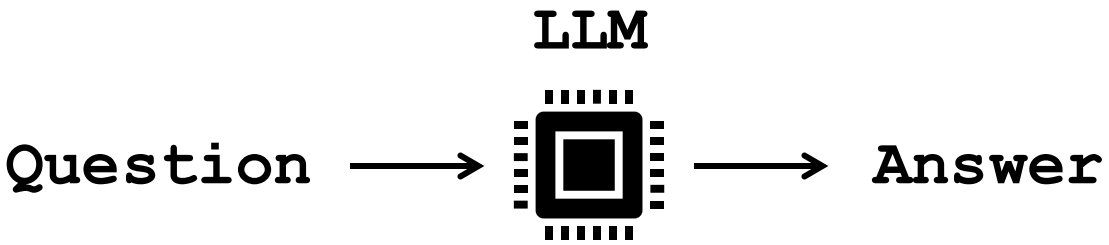
Ali Najafi

Sabancı Üniversitesi

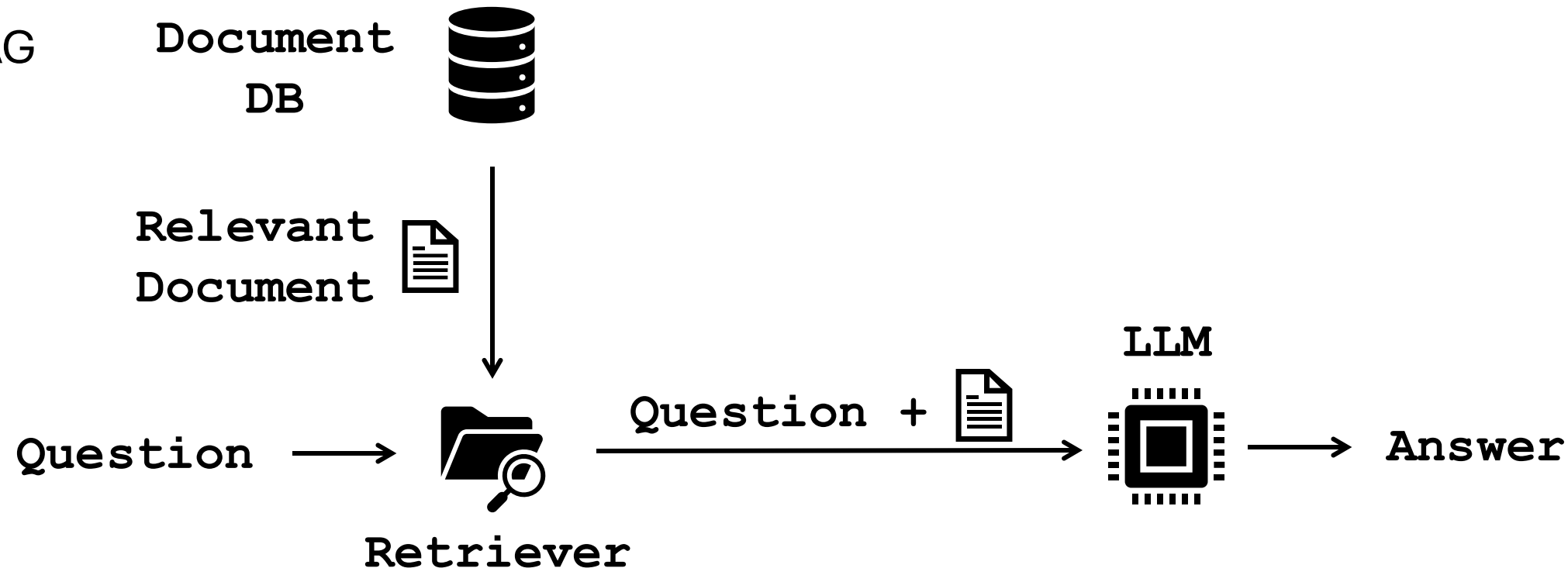
VBYO 2025

What is RAG?

QA in Language Models (LM)



QA using RAG

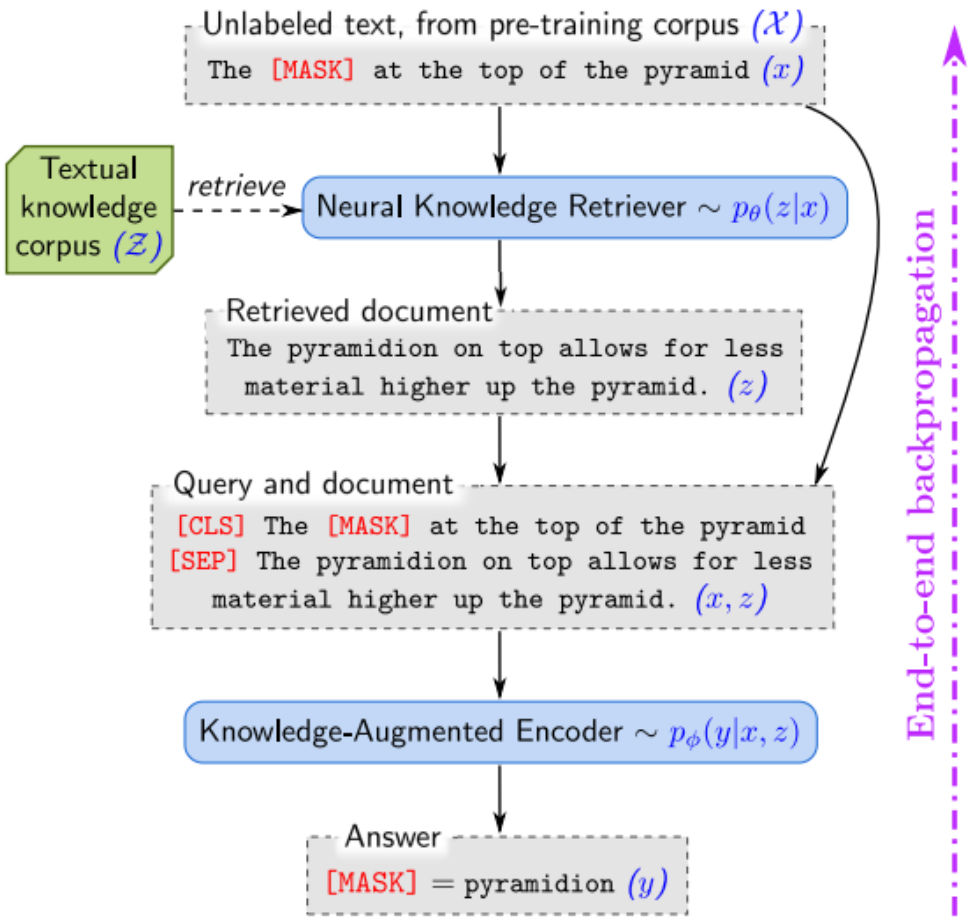


Why RAG?

- Hallucination
- Referencing capabilities
- Up-to-date knowledge
- Private knowledge

REALM: Retrieval-Augmented Language Model Pre-Training

Kelvin Guu [*1] **Kenton Lee** [*1] **Zora Tung** [1] **Panupong Pasupat** [1] **Ming-Wei Chang** [1]



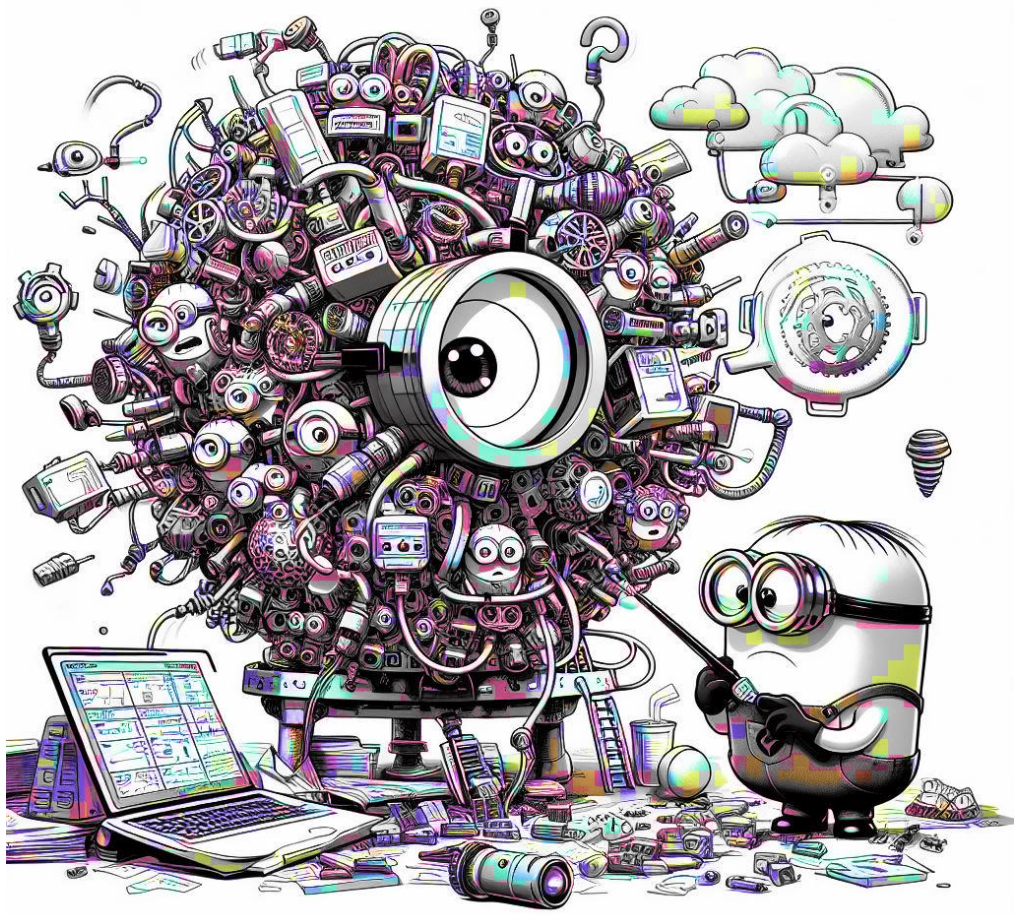
REALM: Retrieval-Augmented Language Model Pre-Training

Kelvin Guu [*1] **Kenton Lee** [*1] **Zora Tung** [1] **Panupong Pasupat** [1] **Ming-Wei Chang** [1]

x:	An equilateral triangle is easily constructed using a straightedge and compass, because 3 is a ___ prime.		
(a) BERT	$p(y = \text{“Fermat”} \mid x)$	$= 1.1 \times 10^{-14}$	(No retrieval.)
(b) REALM	$p(y = \text{“Fermat”} \mid x, z)$	$= 1.0$	(Conditional probability with document z =“257 is … a Fermat prime. Thus a regular polygon with 257 sides is constructible with compass …”)
(c) REALM	$p(y = \text{“Fermat”} \mid x)$	$= 0.129$	(Marginal probability, marginalizing over top 8 retrieved documents.)

Requirements for RAG System

- Encoder language model
- Vector (knowledge) database
- Retrieval procedure
- Generative (decoder) language model



1. Encoder Model

Query: What is the color of the cute small cat?

Documents:
- The funny dog is white.
- The black kitten is cute.
- The grey cat is playful.
- …

?

How can we match the query with the right document?

Encode the query and documents and compute similarities

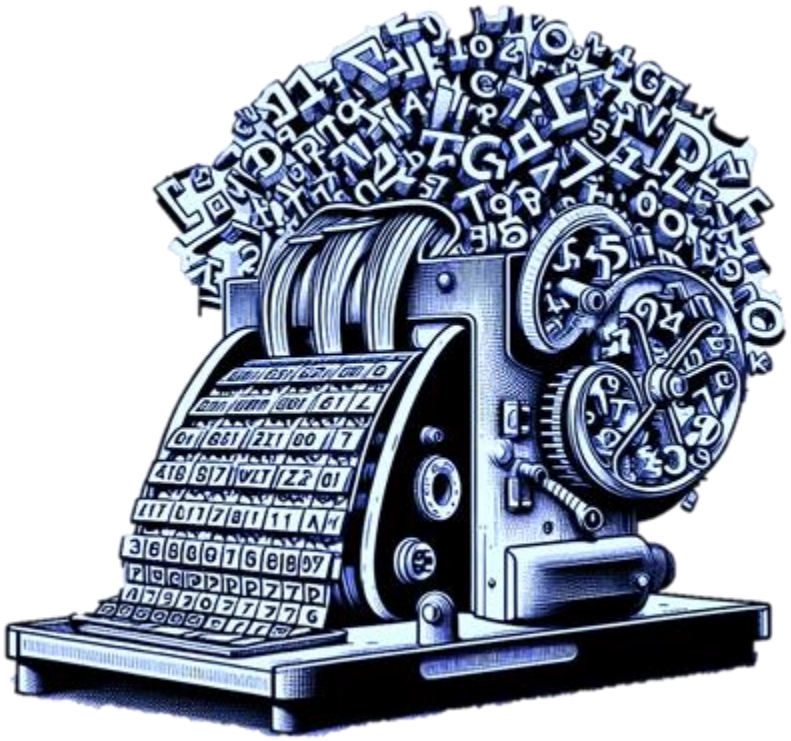
Query → [1, 4, 6, …]

Documents →
[5, 2, 1, …]
[2, 4, 3, …]
[2, 3, 1, …]

1. Encoder model properties

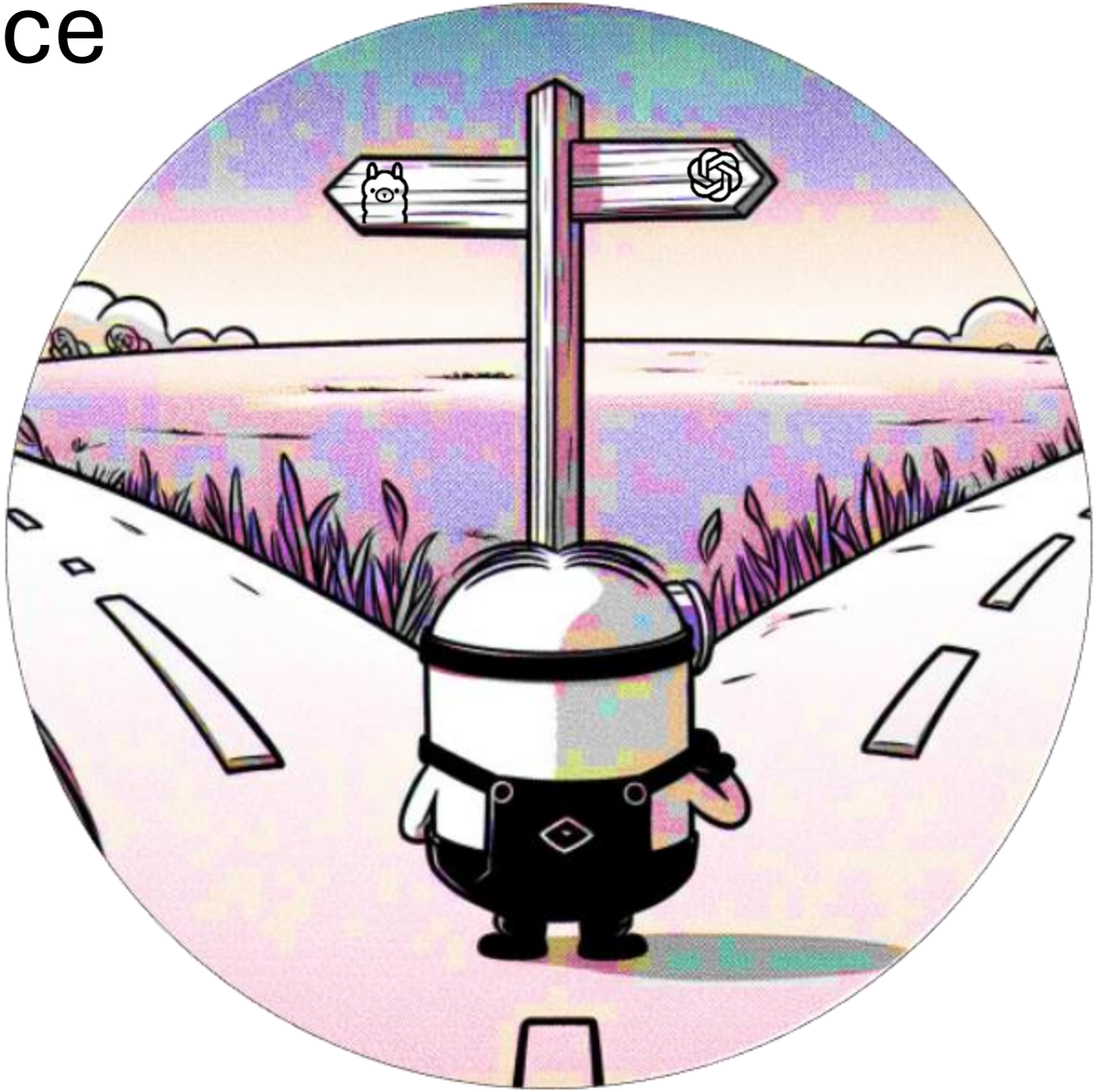
- Provide sequence (not token) encodings
- Encodings capture semantics
- Suitable maximum sequence length

ChatGPT, Google Gemini, open source…



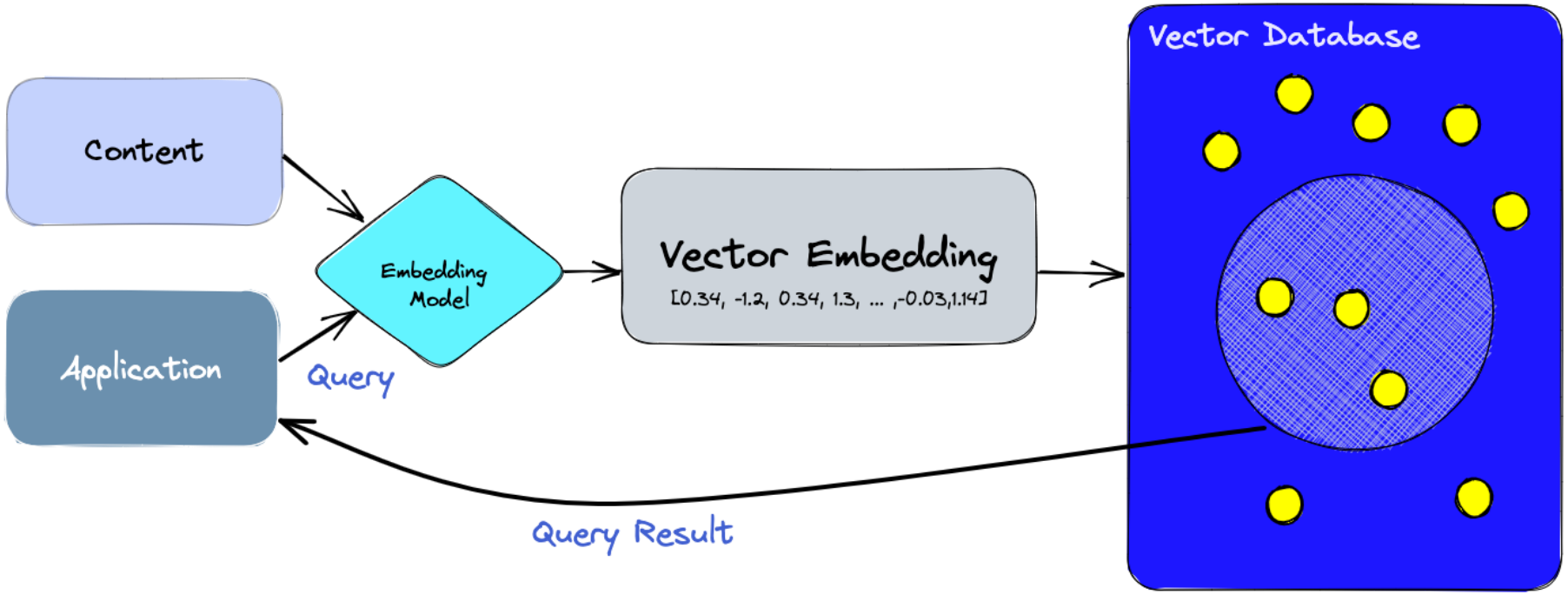
1. Encoder model choice

- Local or cloud?
- Expenses
- Privacy
- Latency



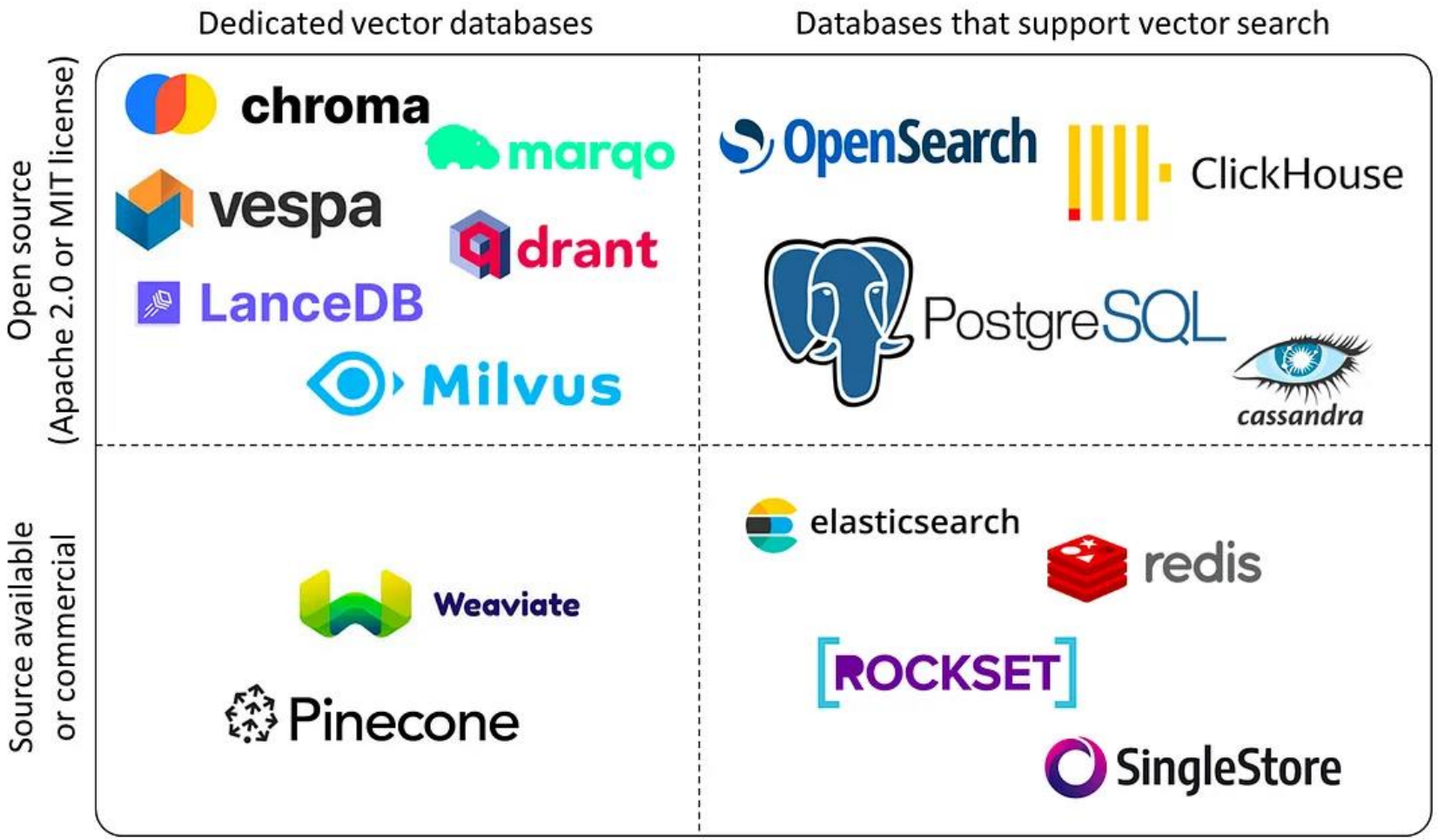
2. Vector Database

- How can we quickly get the most similar document(s)?
- Approximate KNN search



https://www.pinecone.io/learn/vector-database/

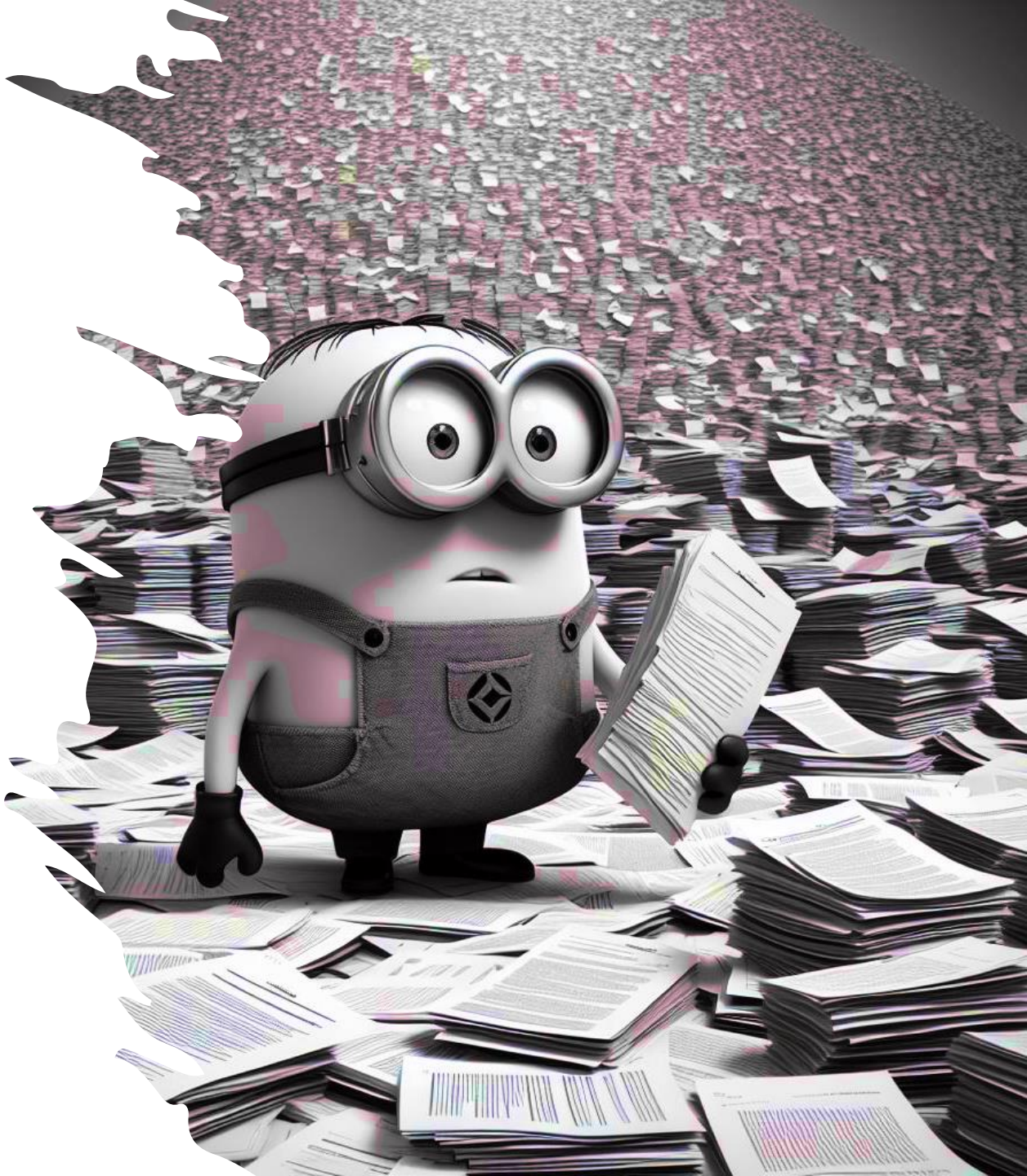
2. Vector Database



Source: https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c

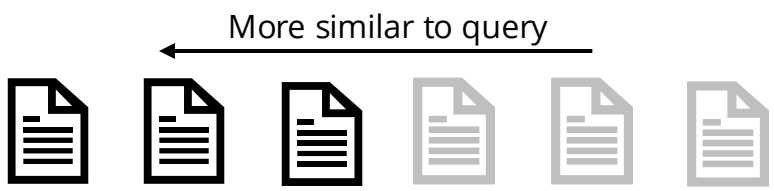
3. Retrieval procedure

- Should we retrieve entire documents, paragraphs, or sentences?
- Limiting factors
  - Semantic content
  - Encoder model
  - Generative model

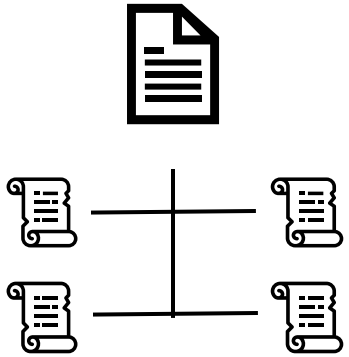


3. Retrieval approaches

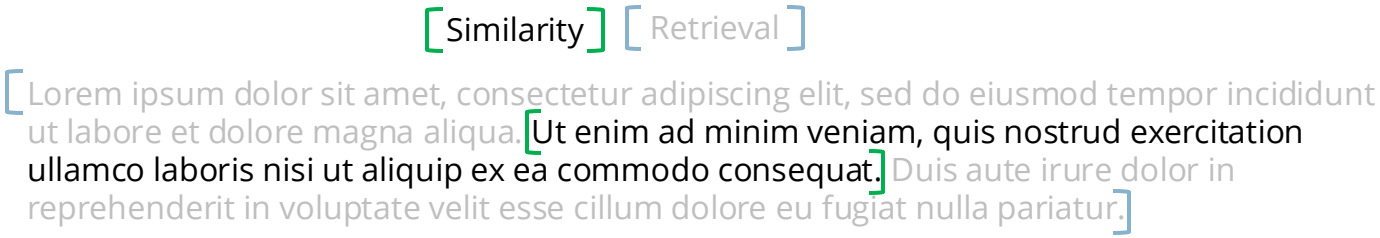
- Get top-k similar documents



- Auto-merging retrieval

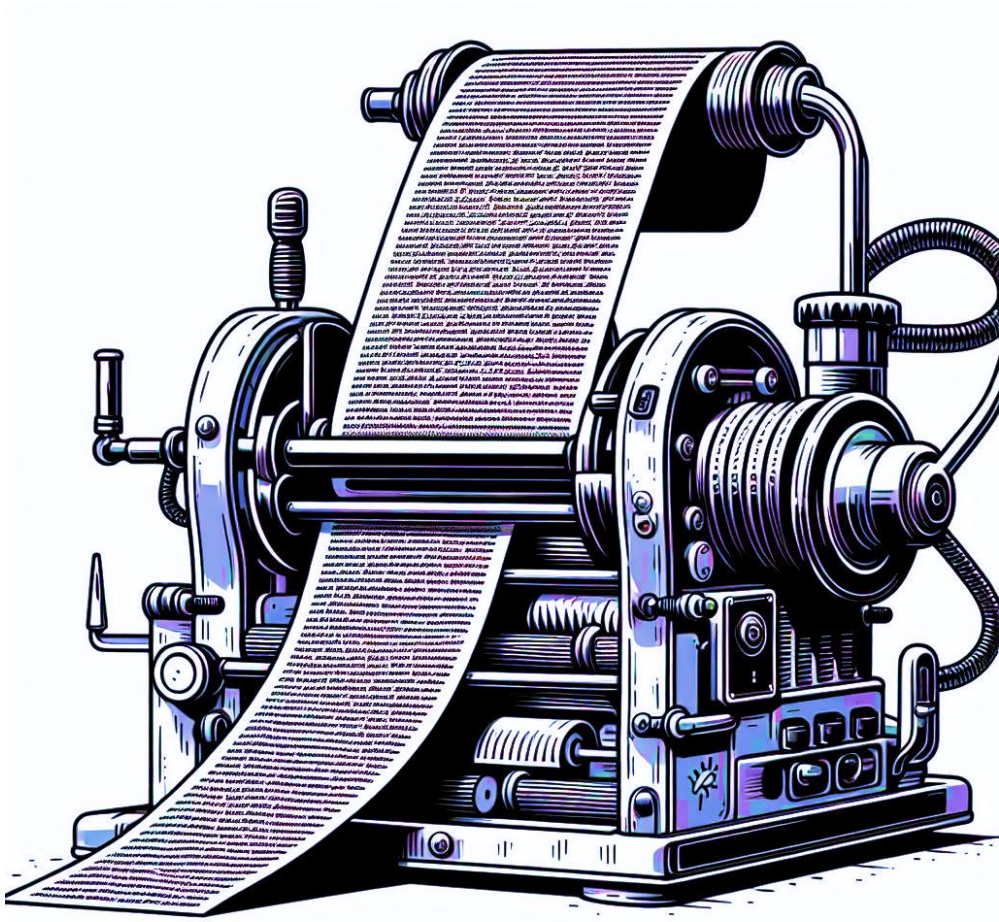


- Sentence window retrieval

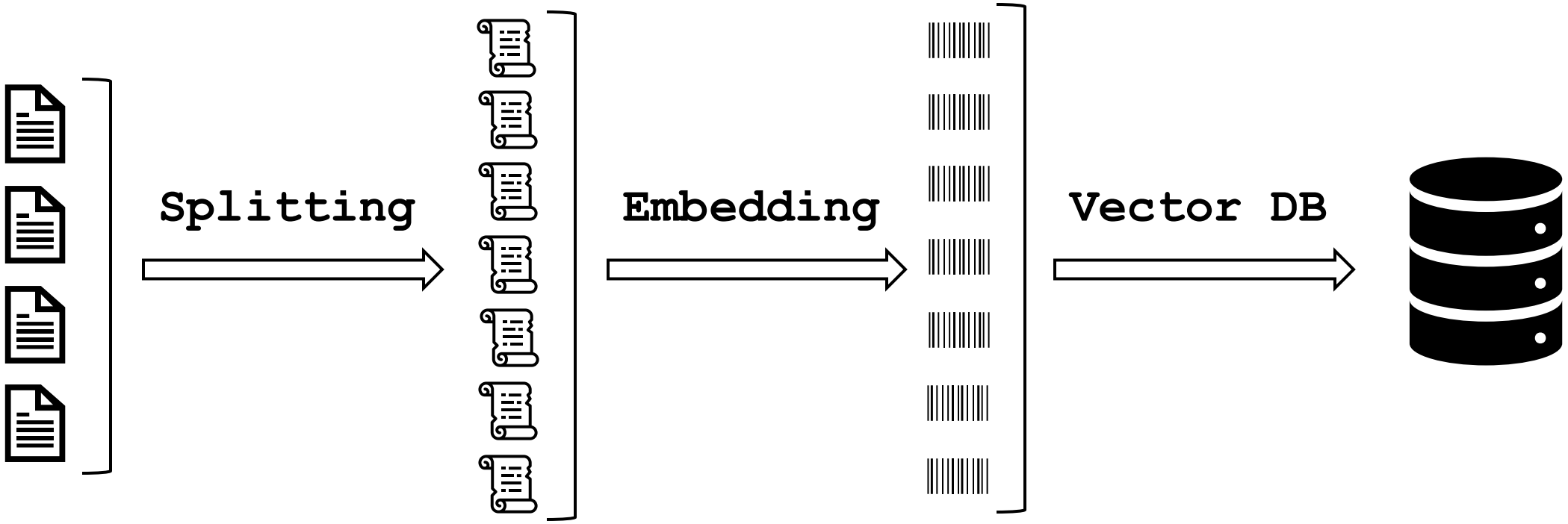


4. Generative model

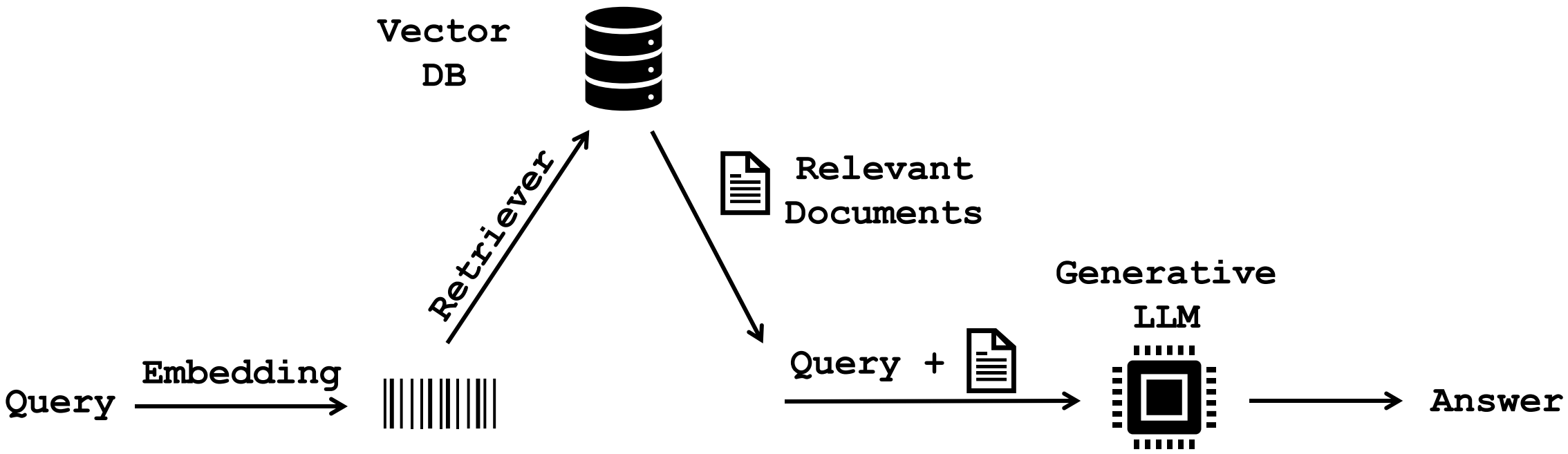
- Local vs cloud
- Context length
- Parameter size
- Latency



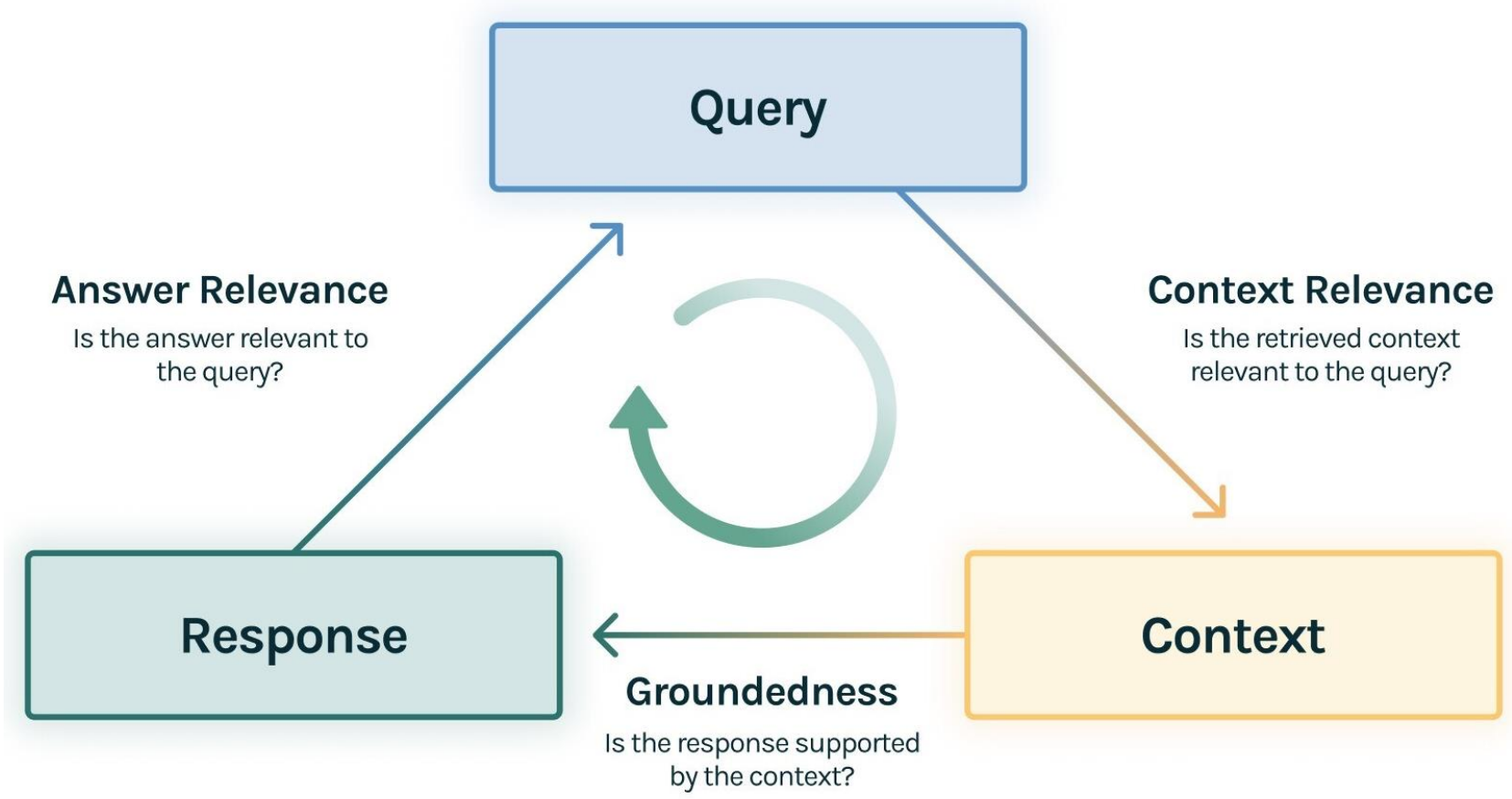
Final pipeline – Creating the system



Final pipeline – Inference



RAG Evaluation



Source: https://www.trulens.org/trulens_eval/core_concepts_rag_triad/

Evaluation – Data Source

- Context relevance
  - Question-Context pairs
- Answer groundedness
  - Answer-Context pairs
- Answer relevance
  - Question-Answer pairs

Evaluation - Metrics

- Lexical matching:
  - Exact matching (EM), BLEU
- Semantic matching:
  - BERTScore, BERT Matching (BEM)
- Auto-Evaluation
  - LLM evaluates LLM
- Human evaluation

Evaluation - Metrics

Evaluating Open-Domain Question Answering in the Era of Large Language Models

Ehsan Kamalloo ♢♣ **Nouha Dziri** ♠ **Charles L. A. Clarke** ♣ **Davood Rafiei** ♢

♢ University of Alberta ♣ University of Waterloo

♠ Allen Institute for Artificial Intelligence

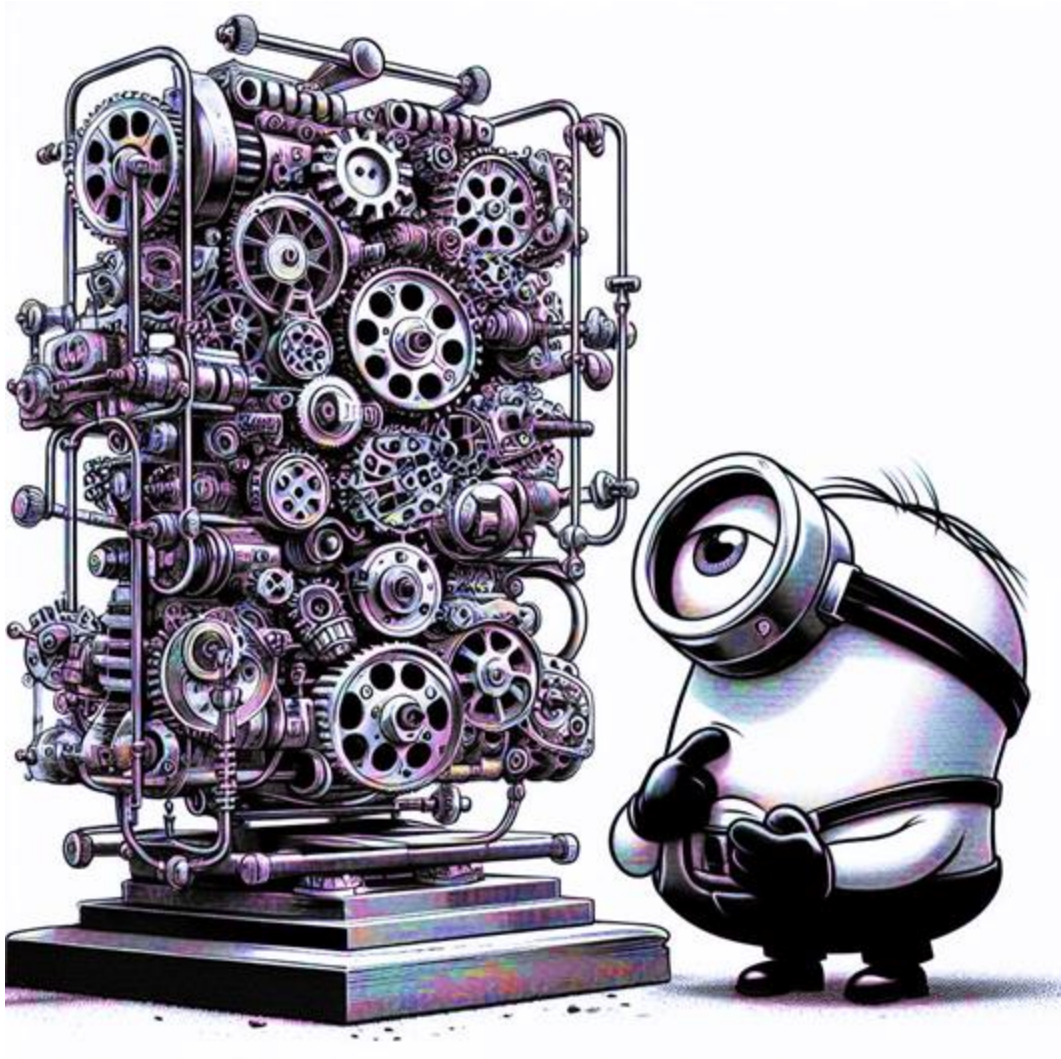
"At this time, there appears to be no substitute for human evaluation."

Problems in evaluation

- Several documents may contain the answer (Context relevance)
- Ground truth answer may not be unique (QA)
- Hallucination (QA Auto-Evaluation)

Advanced RAG

- End-to-end pretraining
  - REALM (Guu et al., 2020)
  - RETRO (Borgeaud et al., 2022)
- Finetuning for RAG
  - Atlas (Izacard et al., 2022)
  - RA-DIT (Lin et al., 2023)
- Reranking



Applications of RAG

- Knowledge engine (ask questions on private data)
- Search augmentation (Bing, Bard)
- Question-Answering chatbots

Source: https://www.databricks.com/glossary/retrieval-augmented-generation-rag

Frameworks for RAG implementation

🦜🔗 LangChain

🦙 LlamaIndex

DSPy

Thank you

- ali.najafi@sabanciuniv.edu
- **www.najafi-ali.com**