



SEMANTIC SEARCH ENGINES OVER BIG DATA USING LLMS

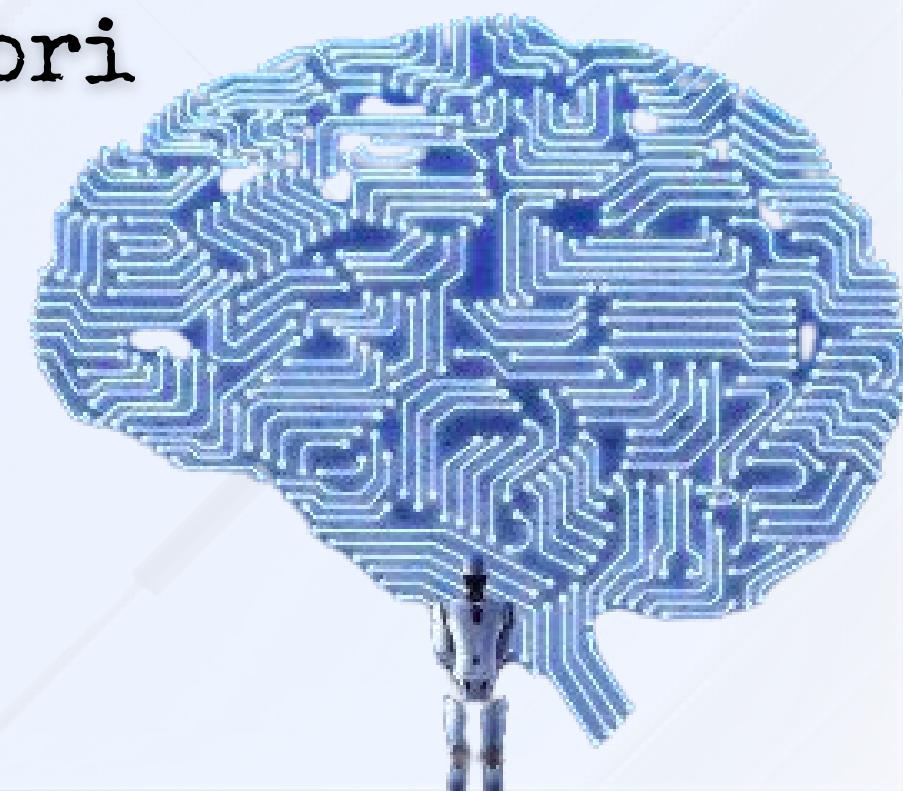
SEARCHING LARGE TEXT CORPORA WITH CONTEXTUAL UNDERSTANDING

Khouloud Ben Younes

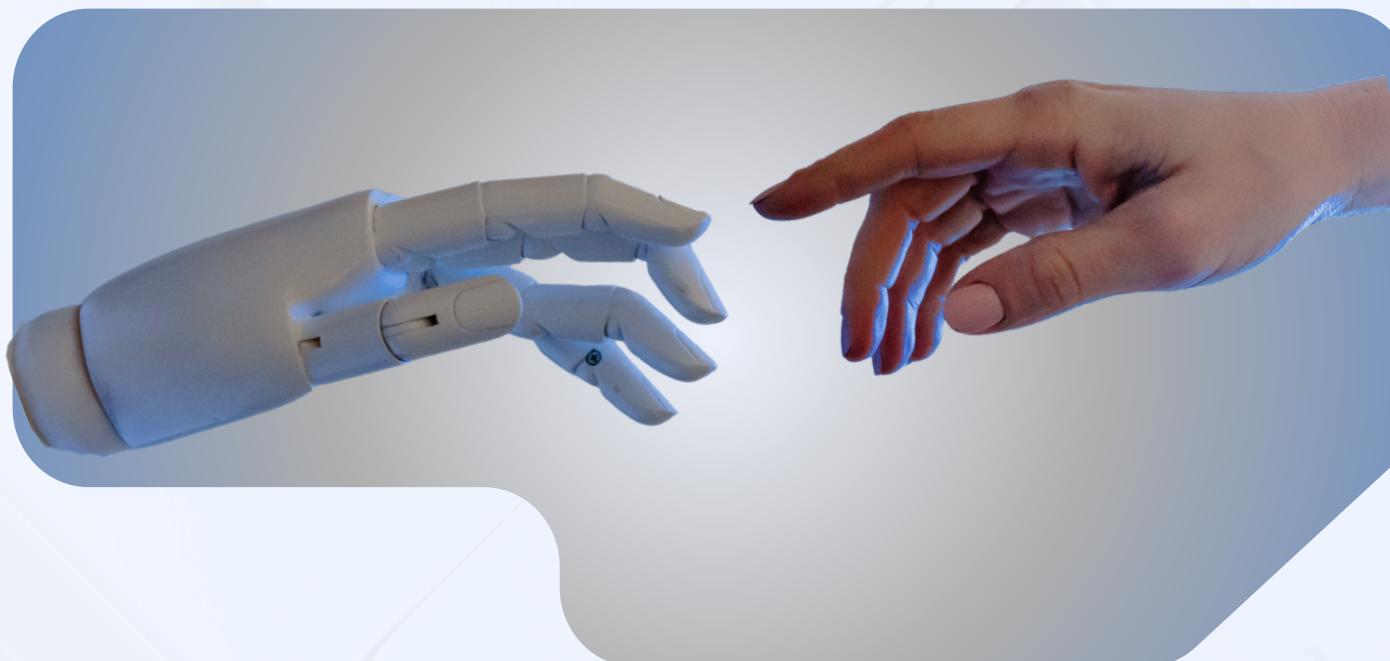
Montaha Ghabri

Evaluated by :
Dr. Manel AbdelKader

2025-2026



AGENDA



01

INTRODUCTION

02

LITERATURE SELECTION

03

BACKGROUND & KEY CONCEPTS

04

REVIEW & ANALYSIS OF EXISTING WORK

05

MAJOR INSIGHTS AND DISCOVERIES

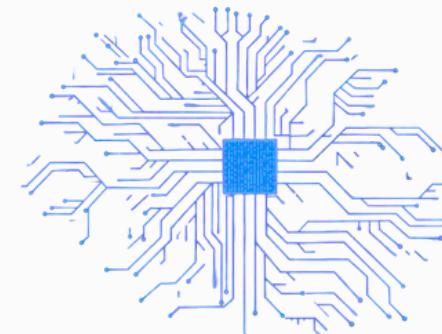
06

CHALLENGES AND FUTURE DIRECTION

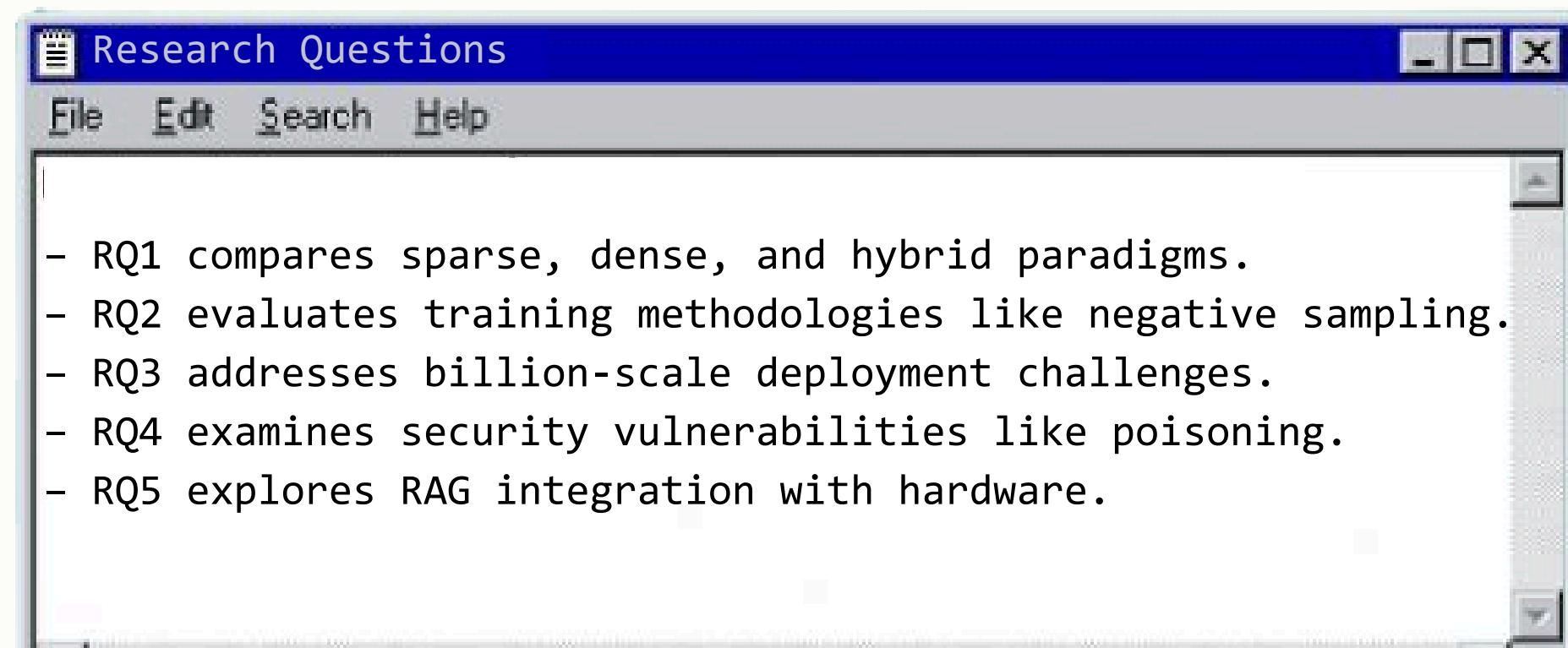
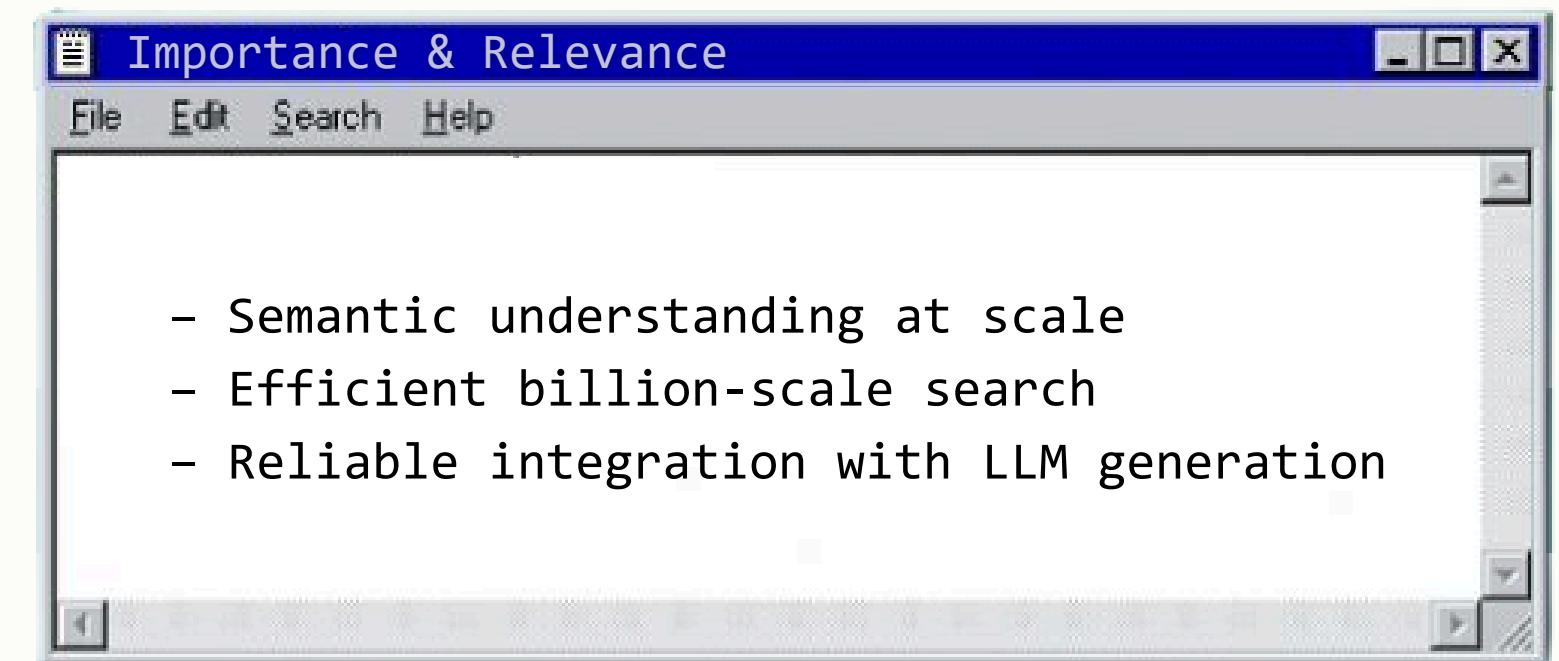
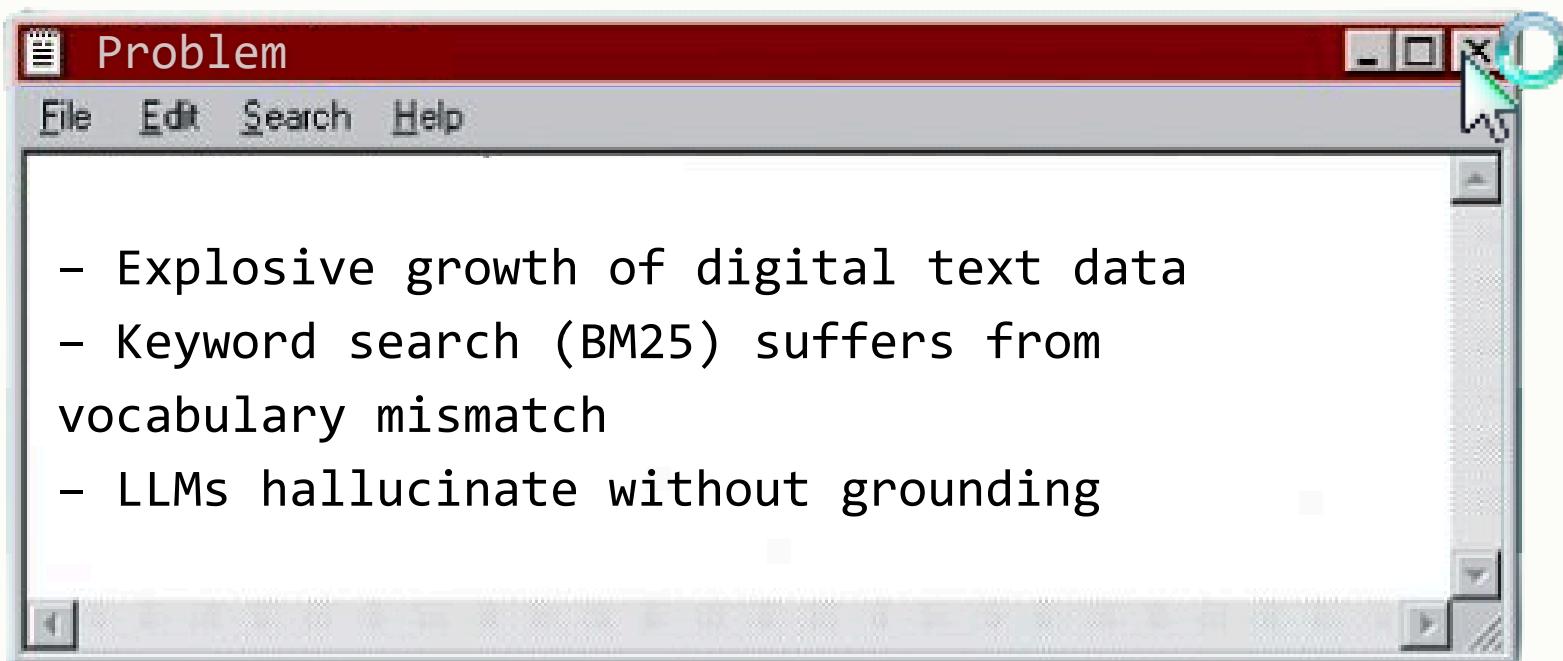
07

CONCLUSION

MOTIVATION



1 MOTIVATION



LITERATURE SELECTION



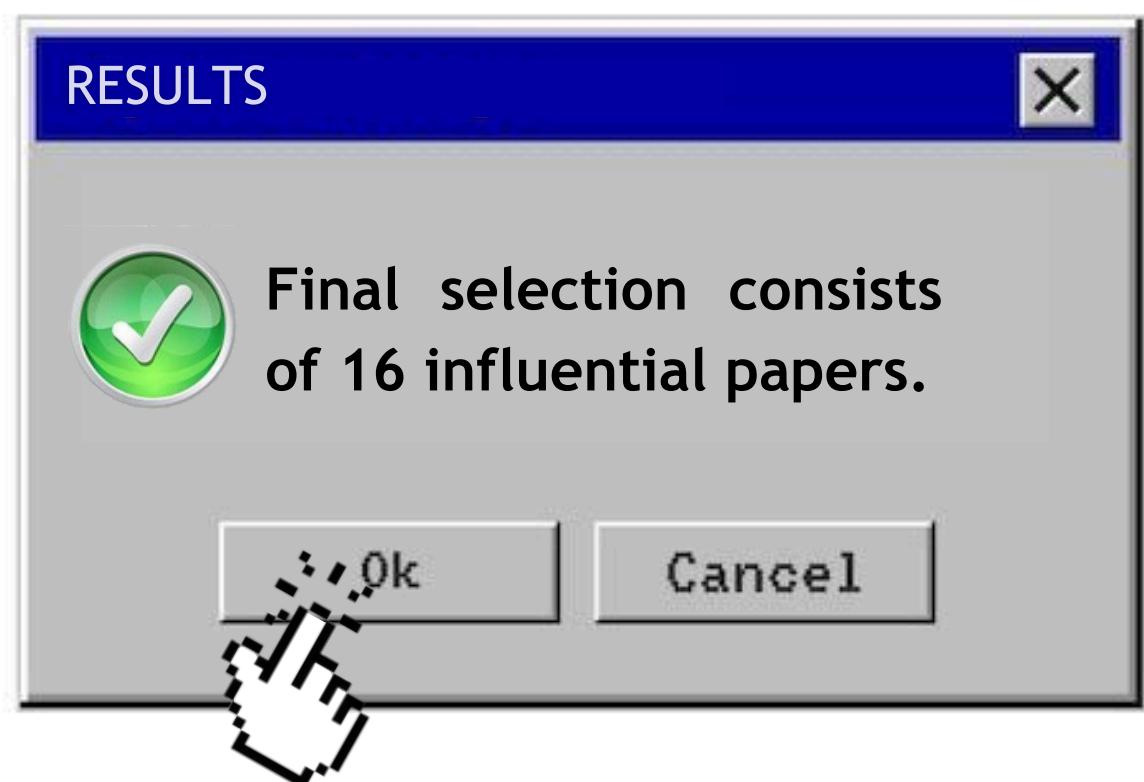
2 LITERATURE SELECTION

Inclusion criteria:

- Top conferences: SIGIR, ACL, EMNLP, NeurIPS, VLDB
- Citation threshold:
 - 100+ citations for 2020-2022 papers
 - 20+ citations for 2023 papers
 - 10+ citations for 2024-2025 papers

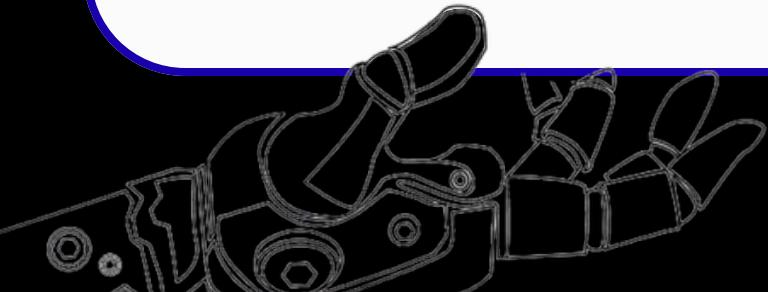
Exclusion criteria:

- Workshop, demo, poster, or short papers
- Non-English publications
- Redundant with earlier seminal works
- Focused only on non-text modalities or unrelated tasks



Title	Authors	Year	Citations	Category	Link	Database	Status	Priority	Notes
Dense Passage Retrieval for Open-Domain Question Answering	Vladimir Karpukhin*, Barlas Oğuz, et al.	2020	5695	Dense Retrieval Foundations	https://arxiv.org/pdf/2004.0490				
SIMLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval	Liang Wang, Nan Yang, Xiaolo	2023	148	Dense Retrieval Foundations	https://aclanthology.org/2023				

BACKGROUND & KEY CONCEPTS



3 Background & Key Concepts

Definitions

-  **Embeddings:** ▾
Dense vector representations (768-1536 dimensions)
-  **Dense Retrieval** ▾
-  **Contrastive Learning** ▾
-  **Hard negatives** ▾
-  **Vector Quantization (VQ)** ▾
-  **Suspend**

 Start

3 Background & Key Concepts

The image shows a software application window with a dark blue header bar at the top. In the top-left corner of the header, there is a small blue icon consisting of three dots connected by lines. The main area of the window has a light gray background. On the left side, there is a vertical sidebar with a dark gray header labeled "Definitions". Inside the sidebar, there is a list of items, each with an icon and a label:

- Embeddings:** A dropdown menu is open next to this item, showing the text "Dual encoders with ANN search (FAISS, HNSW)".
- Dense Retrieval**
- Contrastive Learning**
- Hard negatives**
- Vector Quantization (VQ)**
- Suspend**

At the bottom of the sidebar is a button labeled "Start" with a small icon to its left. The main panel of the application is mostly empty, except for a small blue arrow icon in the top right corner.

3 Background & Key Concepts

Definitions

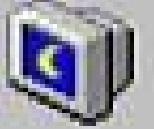
- Embeddings:
- Dense Retrieval
- Contrastive Learning
- Hard negatives
- Vector Quantization (VQ)
- Suspend

Start

Training objective pulling positive query-passage pairs close while pushing negative pairs apart

3 Background & Key Concepts

Definitions

-  **Embeddings:** ↗
-  **Dense Retrieval** ↗
-  **Contrastive Learning** ↗
-  **Hard negatives** ↘
 - semantically similar but incorrect passages that improve discrimination
-  **Vector Quantization (VQ)** ↗
-  **Suspend**

Start

3 Background & Key Concepts

Definitions

- Embeddings:
- Dense Retrieval
- Contrastive Learning
- Hard negatives
- Vector Quantization (VQ)
 - Compression methods replacing full-precision vectors with codes
- Suspend

Start

REVIEW OF EXISTING WORK

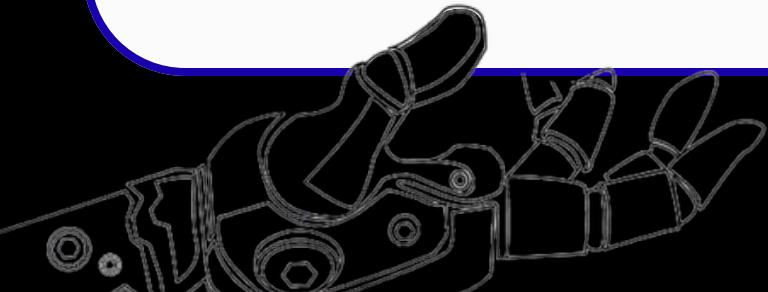


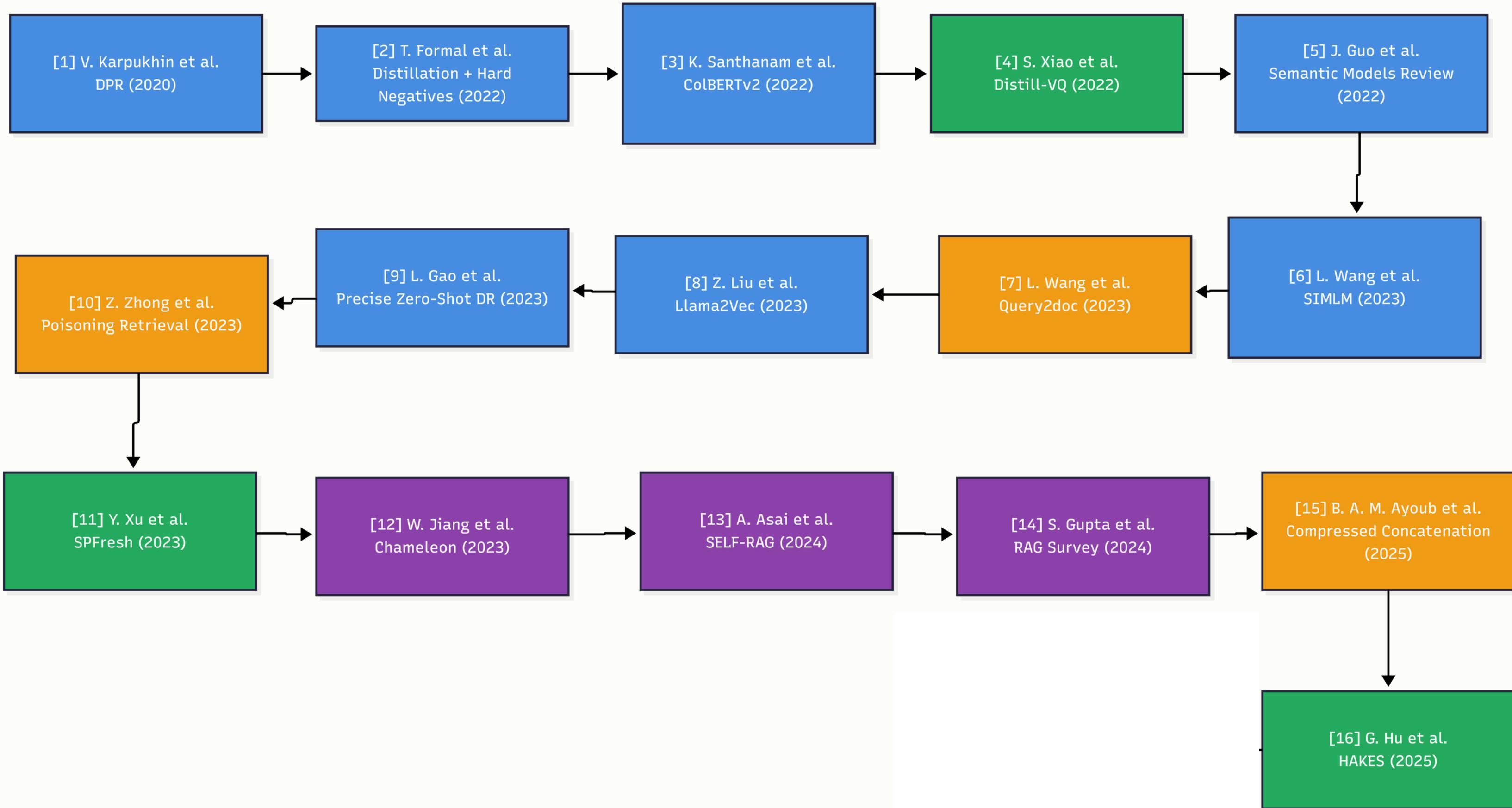
You are offmark

Try:

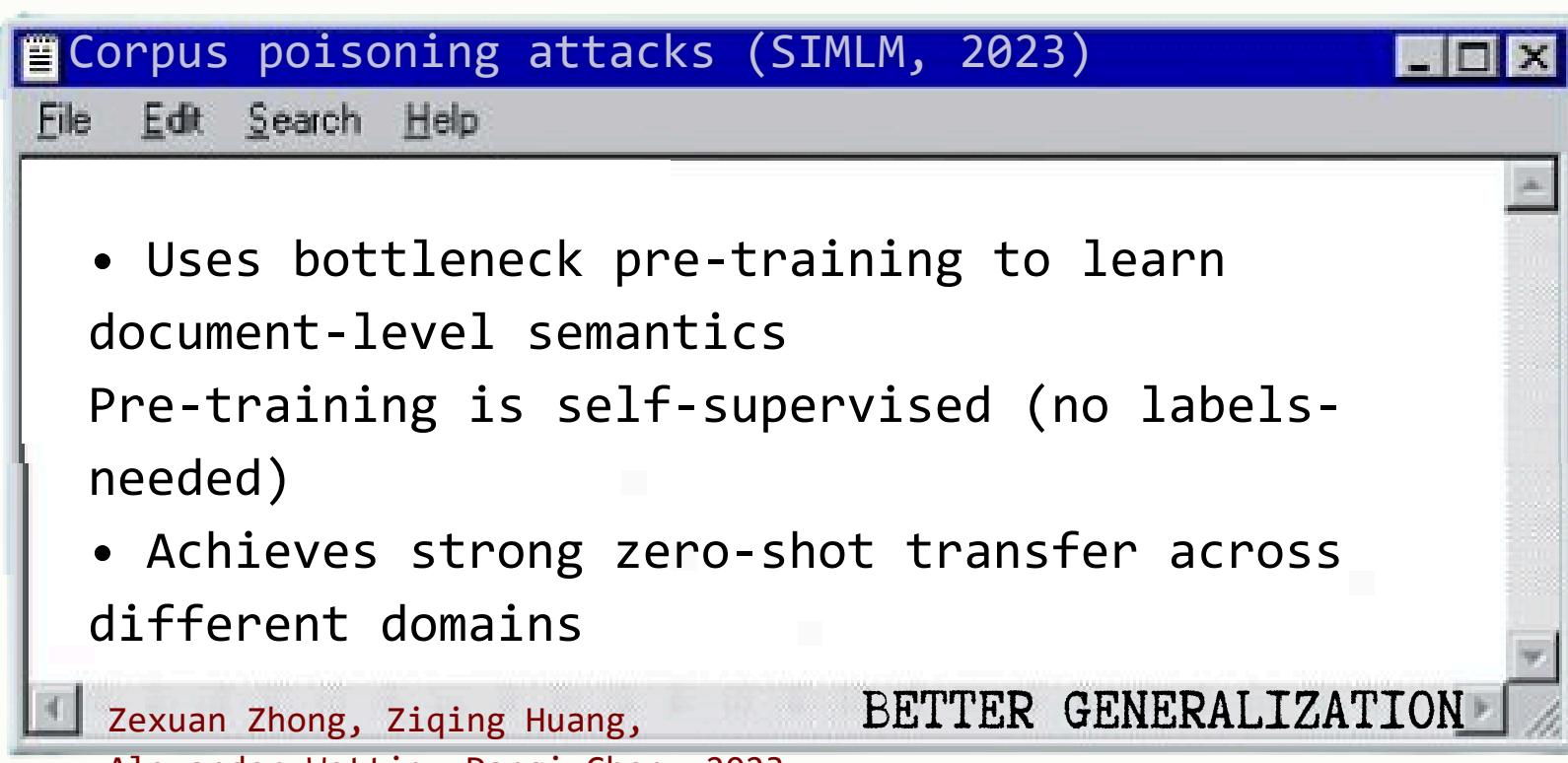
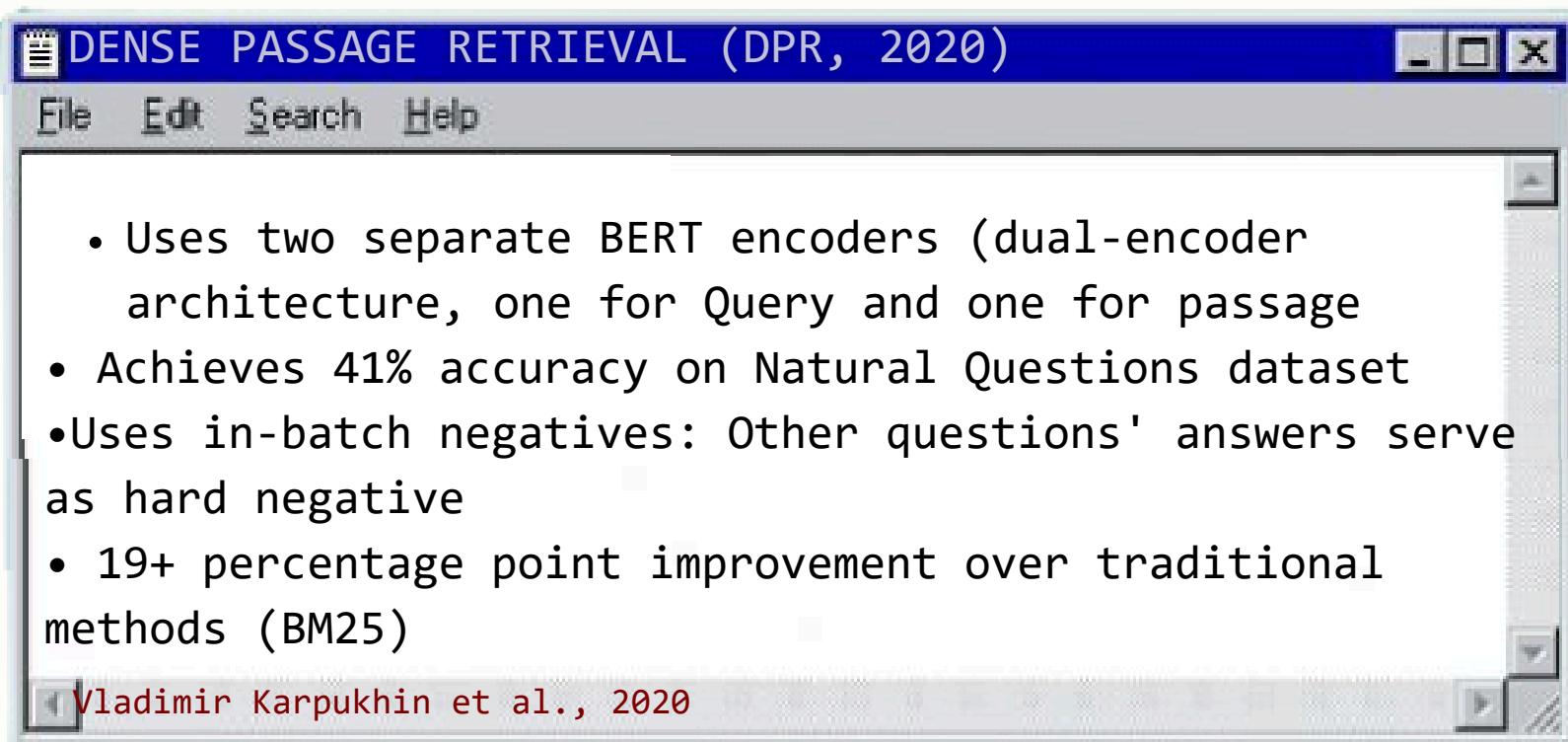
Reading recent papers.

FOUR RESEARCH CATEGORIES

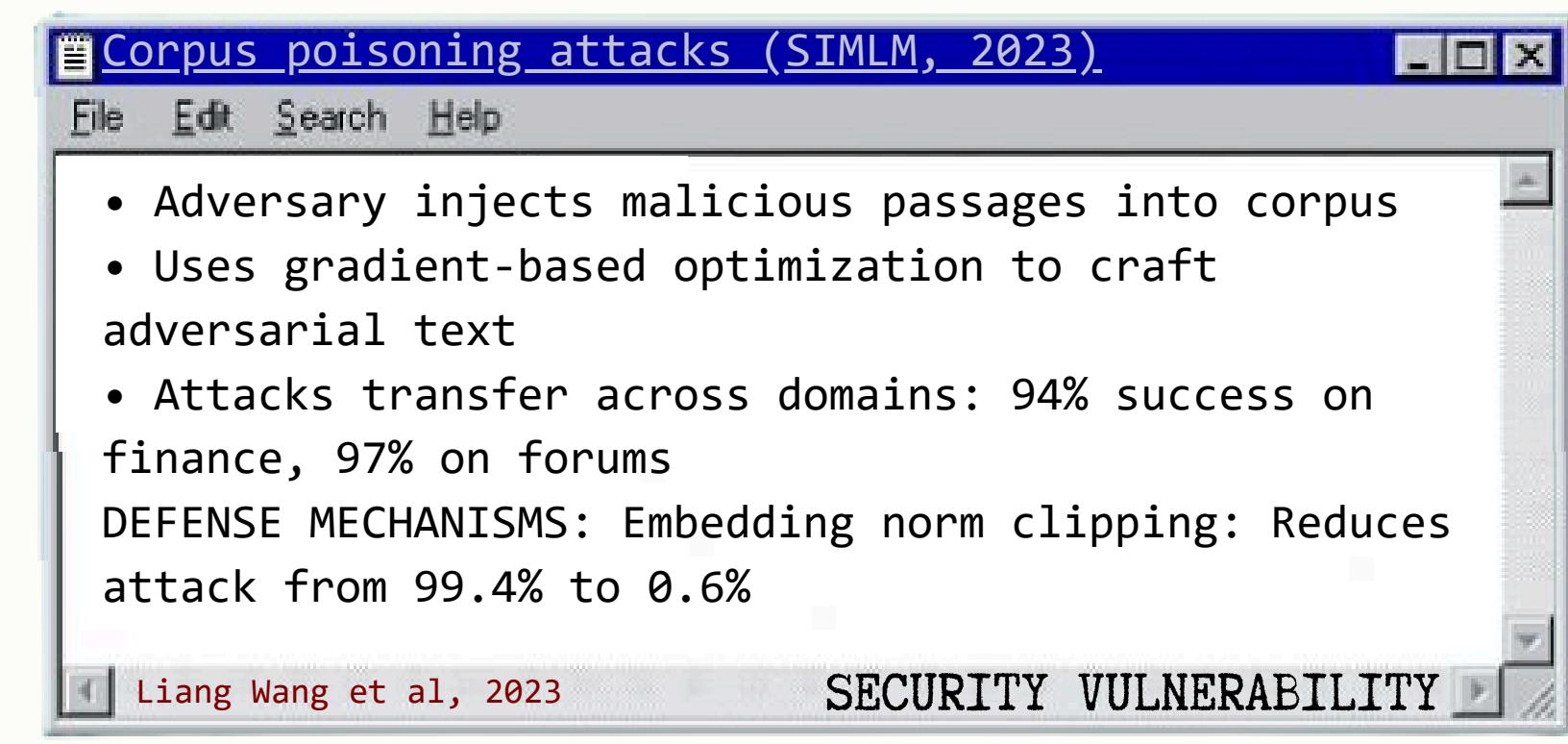




1- DENSE RETRIEVAL FOUNDATIONS



BETTER GENERALIZATION



SECURITY VULNERABILITY

2- SCALABILITY SOLUTIONS FOR PRODUCTION

HAKES: Scalable Vector Database for Embedding_Search Service

- ✖ Traditional databases couple storage and computation!
Can't scale read and write operations independently!

TWO-STAGE ARCHITECTURE:

STAGE 1 - Fast Filtering: Quickly narrows down to candidate documents

STAGE 2 - Precise Refinement: Exact similarity computation on candidates

Handles concurrent read-write workloads efficiently

Guoyu Hu et al. , 2025

SPFresh- INCREMENTAL INDEX UPDATES

- ✖ Traditional indices require full rebuilds when data changes !

- Lightweight Incremental Rebalancing (LIRE) approach
- When new vectors arrive, only rebalance overloaded partitions
- 2-5x faster updates on billion-vector indices

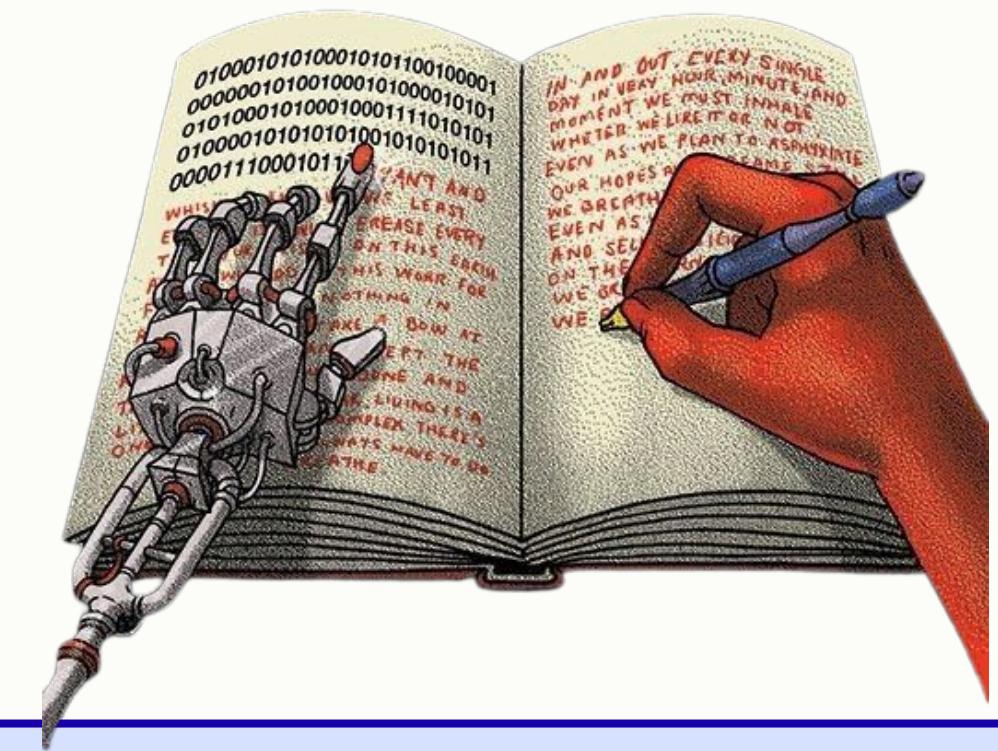
Yuming Xu et al., 2023

Distill-VQ: Learning Retrieval Oriented VQ by distilling knowledge

- ✖ 1. Product Quantization (PQ) compresses vectors to save memory.
- 2. Reconstructed vectors ≠ good ranking preservation

- Uses knowledge distillation from a frozen teacher retriever
- Teacher provides soft relevance judgments for query-document pairs
- Student (compressed model) learns to preserve teacher's ranking
- No labeled data needed - teacher provides supervision, 2-5% MRR and recall improvement vs standard PQ

Shitao Xiao et al. , 2022



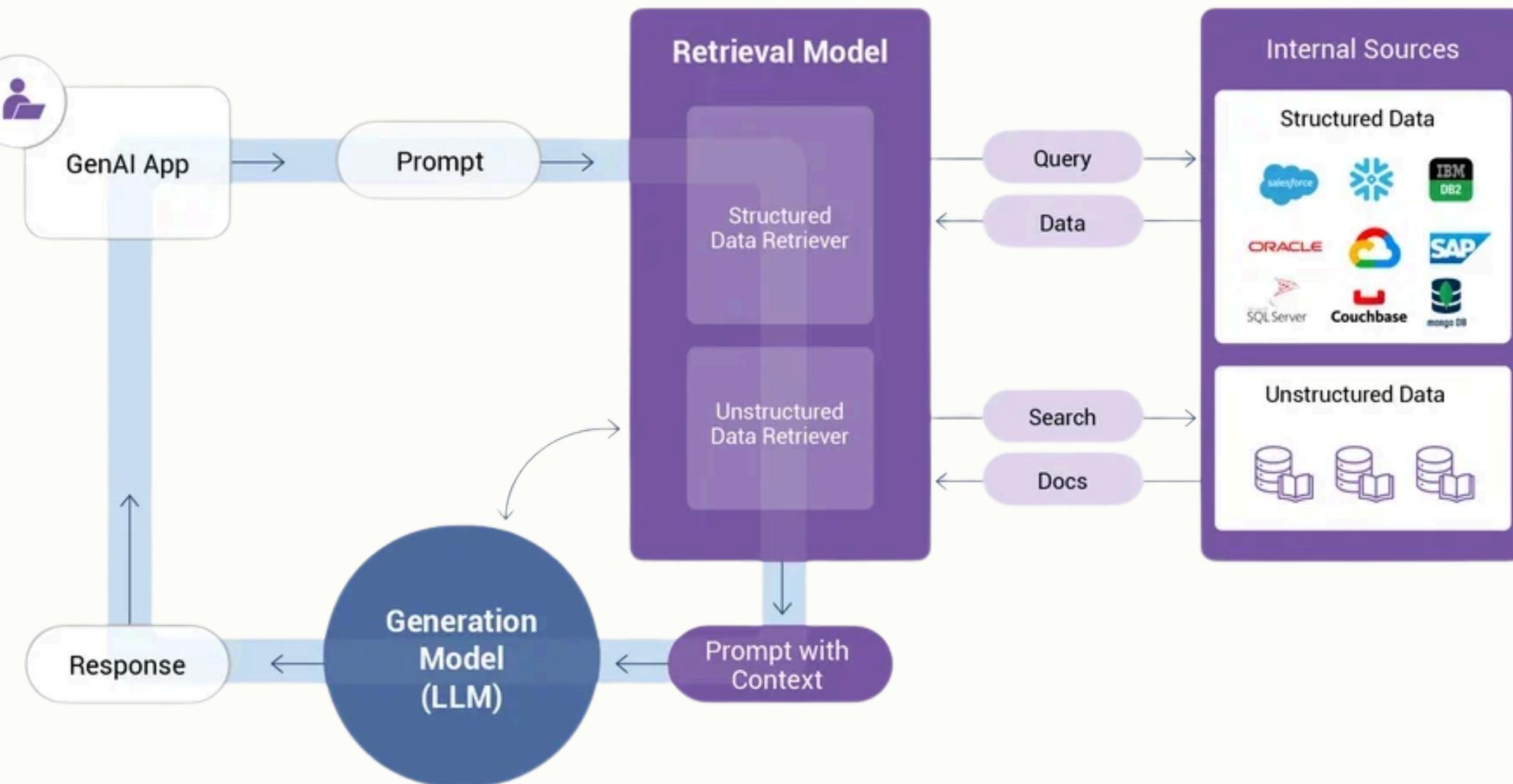
3- RAG ARCHITECTURES

WHY RAG?

- LLMs can generate plausible but incorrect information
- Knowledge becomes outdated (training data cutoff)
- RAG grounds generation in retrieved factual documents
- Enables knowledge updates without retraining

Shailja Gupta, Rajesh Ranjan, Surya Narayan Singh, 2024

COMBINE RETRIEVAL WITH LLM GENERATION TO REDUCE HALLUCINATIONS



3- RAG ARCHITECTURES

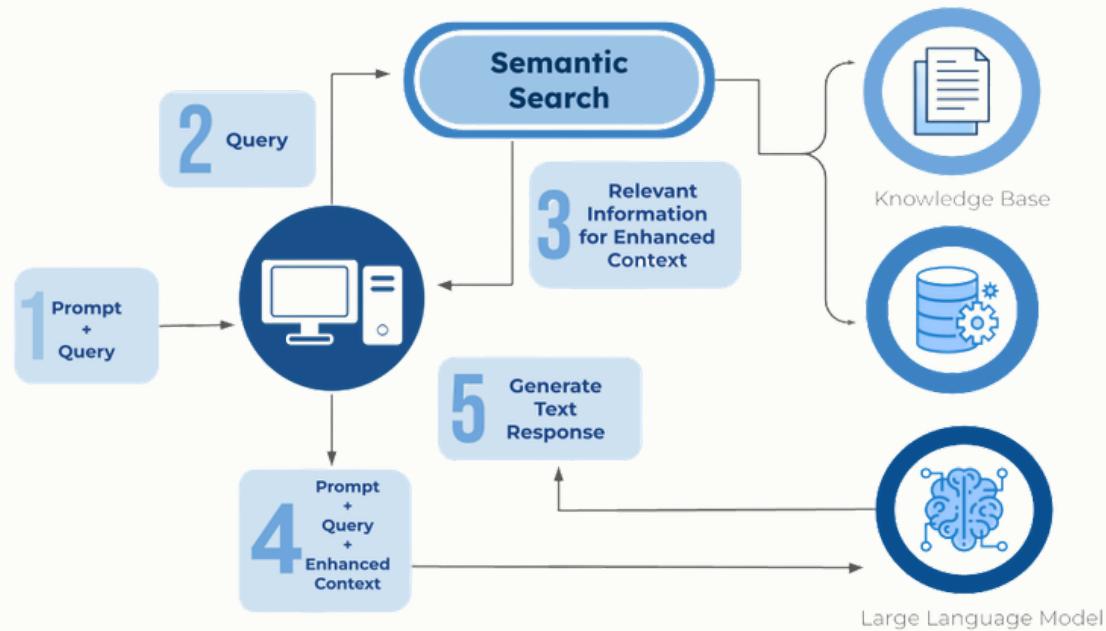
Combine retrieval with LLM generation to reduce hallucinations

SELF-RAG - INTELLIGENT RETRIEVAL

- Model generates special control tokens alongside content:
 - [Retrieve]: Decides when to trigger retrieval system
 - [Relevant]/[Irrelevant]: Assesses if retrieved docs relate to query
 - [Supported]/[Not Supported]: Checks if output is grounded in evidence
 - [Utility]: Evaluates overall output quality
- Critic model automatically annotates training data
- Main model learns to predict these tokens via next-token prediction

Reduces unnecessary retrievals by 50%

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, Hannaneh Hajishirzi, 2024



QUERY2DOC

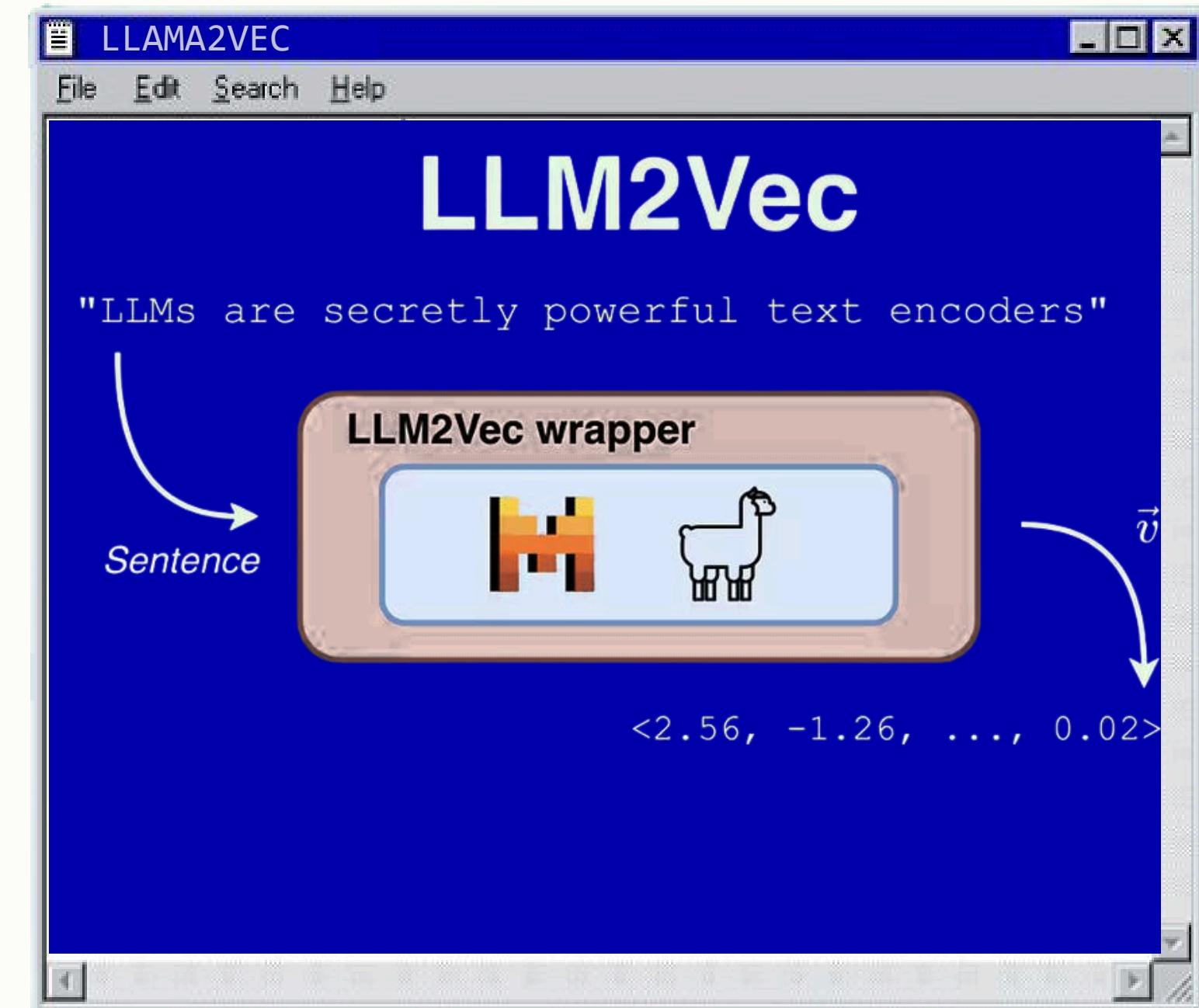
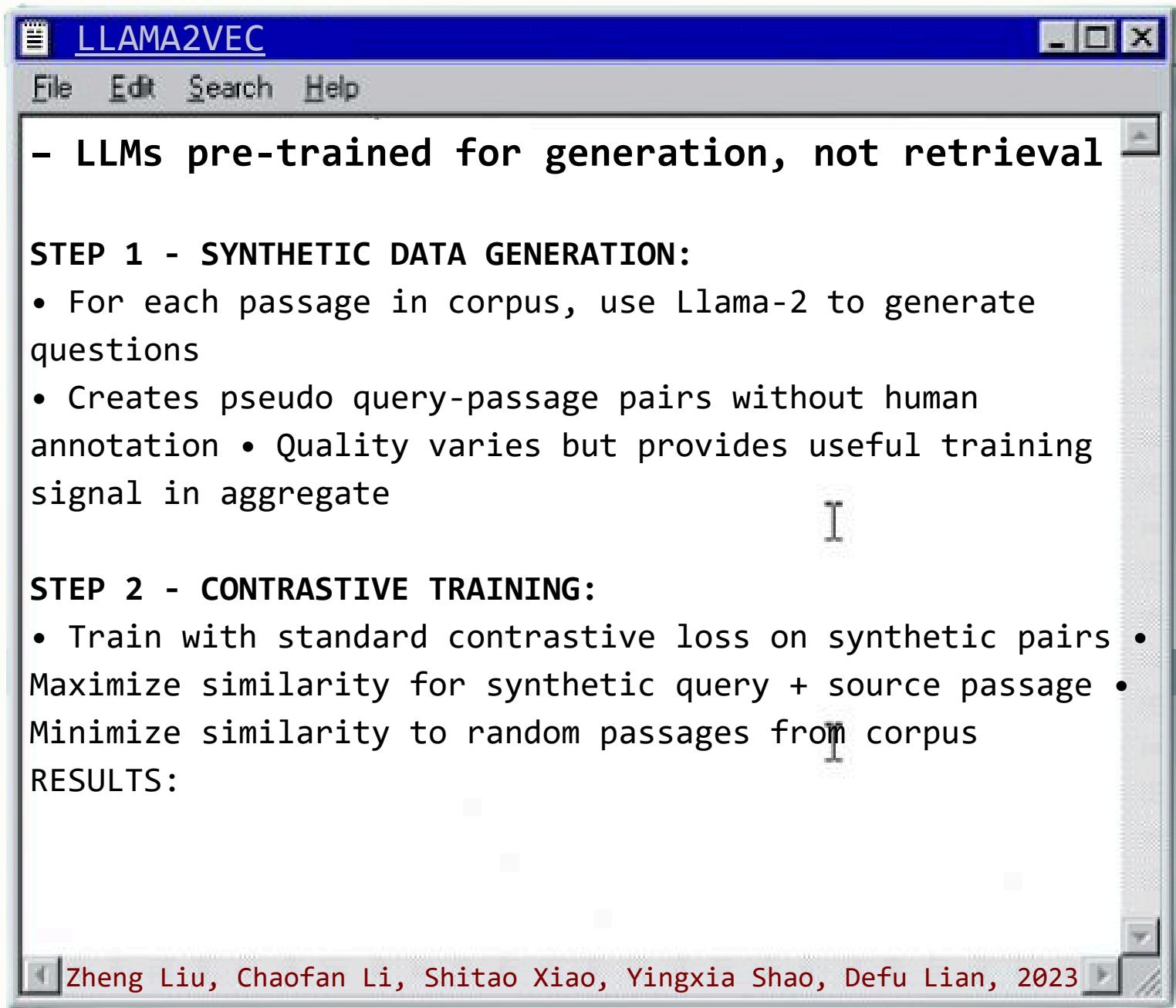
Queries are short and may lack vocabulary of relevant documents !

- Uses LLM to generate hypothetical relevant document for query
- Concatenates generated text with original query
- No training or labeled data needed for target task
- Compatible with both BM25 (sparse) and dense retrievers
- Up to 15% improvement on BM25 (zero-shot)

Liang Wang, Nan Yang, Furu Wei, 2023

3- RAG ARCHITECTURES

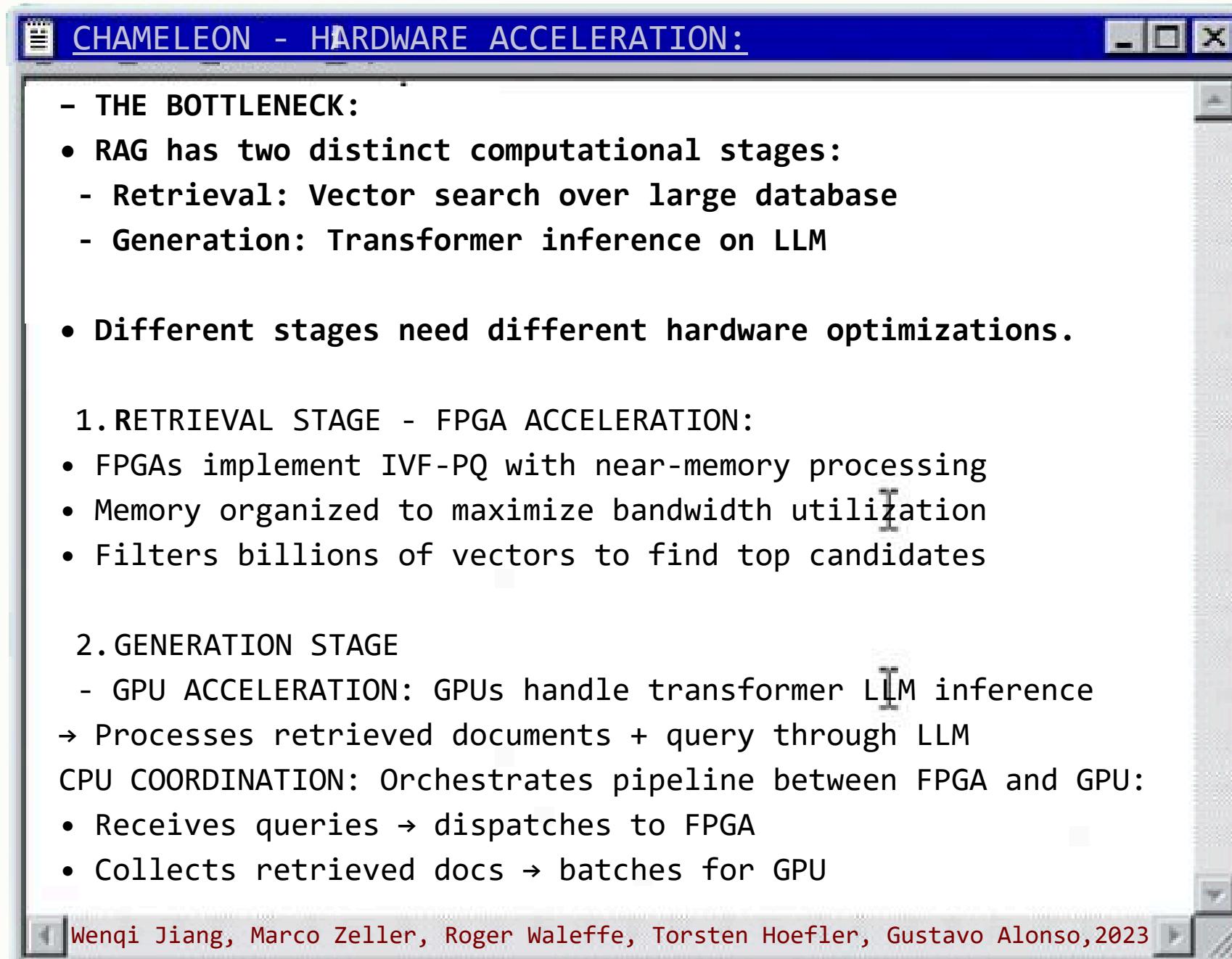
Combine retrieval with LLM generation to reduce hallucinations



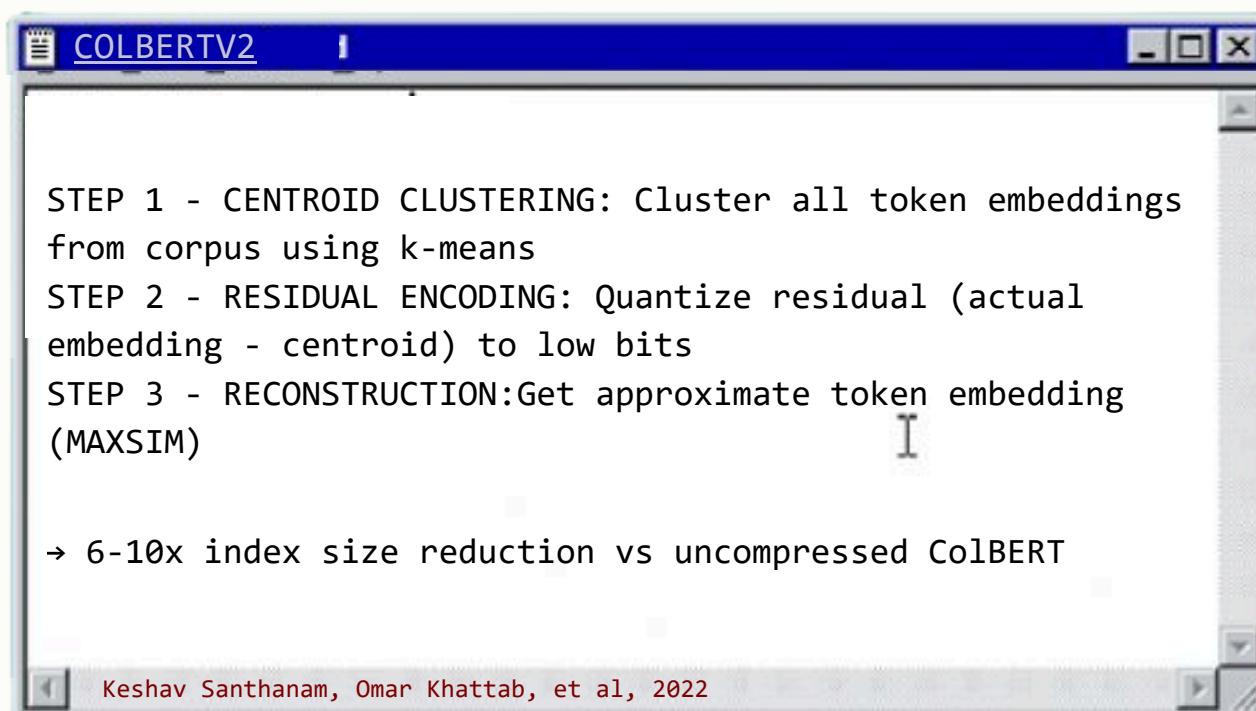
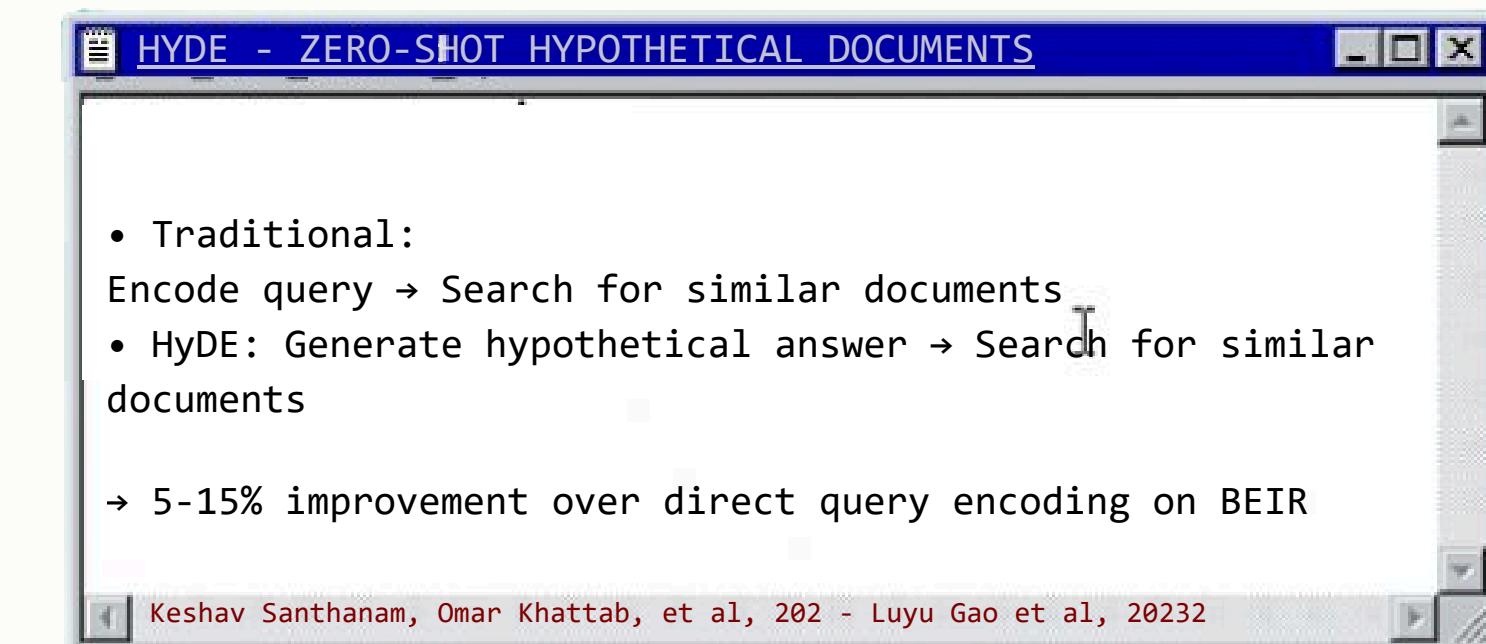
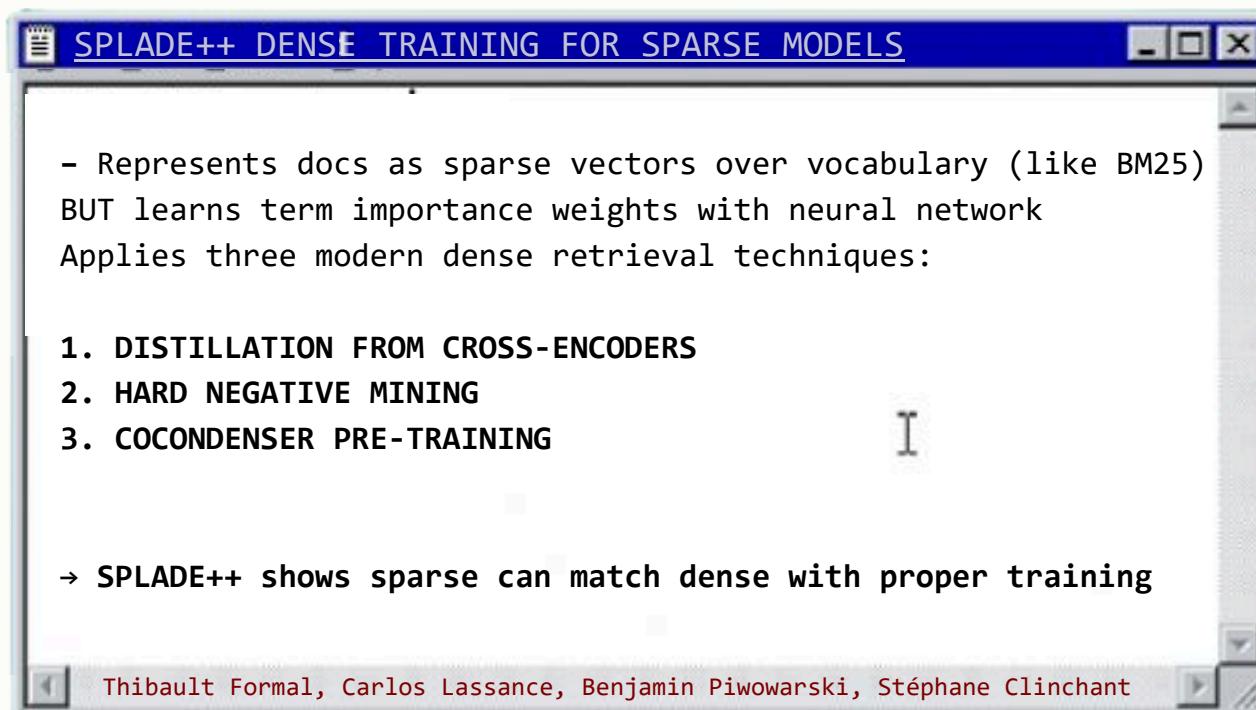
- Matches supervised dense retrievers on BEIR (out-of-domain)
- Enables deployment in domains without labeled retrieval data

3- RAG ARCHITECTURES

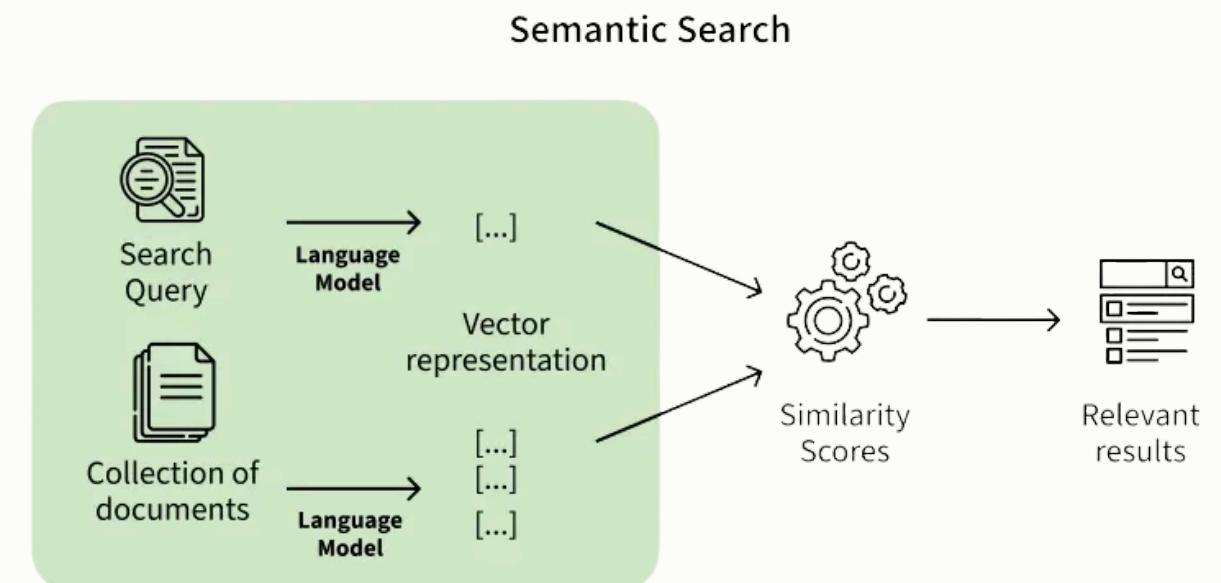
Combine retrieval with LLM generation to reduce hallucinations



4- HYBRID & ADVANCED METHODS



4- HYBRID & ADVANCED METHODS



SPLADE++ DENSE TRAINING FOR SPARSE MODELS

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, Stéphane Clinchant

Represents docs as sparse vectors over vocabulary (like BM25)

BUT learns term importance weights with neural network

Applies three modern dense retrieval techniques:

1. DISTILLATION FROM CROSS-ENCODERS
2. HARD NEGATIVE MINING
3. COCONDENSER PRE-TRAINING

→ SPLADE++ shows sparse can match dense with proper training

COLBERTV2

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, Matei Zaharia, 2022

STEP 1 - CENTROID CLUSTERING: Cluster all token embeddings from corpus using k-means

STEP 2 - RESIDUAL ENCODING: Quantize residual (actual embedding - centroid) to low bits

STEP 3 - RECONSTRUCTION: Get approximate token embedding (MAXSIM)

→ 6-10x index size reduction vs uncompressed ColBERT

HYDE - ZERO-SHOT HYPOTHETICAL DOCUMENTS

Luyu Gao, Xueguang Ma, Jimmy Lin, Jamie Callan, 2023

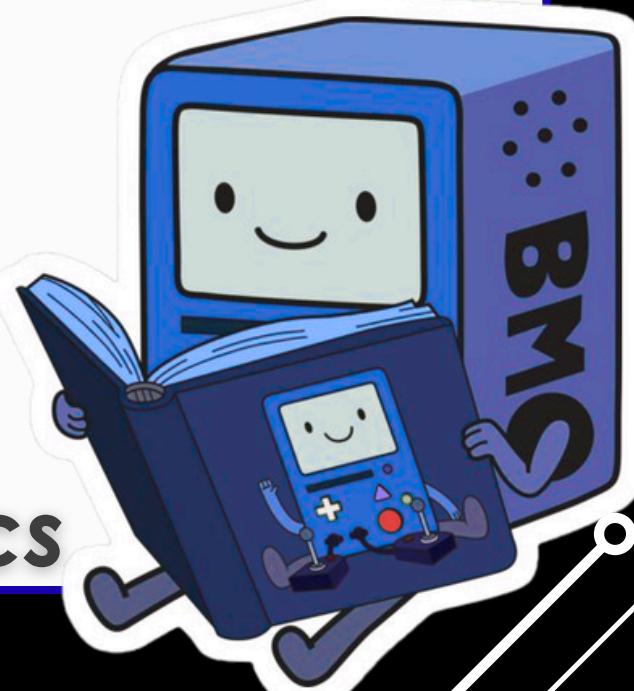
Traditional: Encode query → Search for similar documents

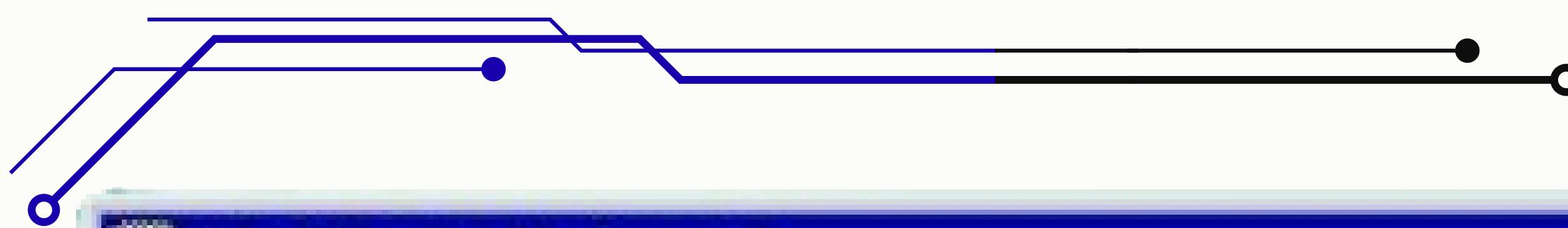
- HyDE: Generate hypothetical answer → Search for similar documents

→ 5-15% improvement over direct query encoding on BEIR

MAJOR INSIGHTS AND DISCOVERIES

Recent Directions and Emerging Topics





INSIGHTS

TRAINING METHODOLOGY MATTERS MOST

- Simple architectures with smart training beat complex designs
 - DPR's dual-encoder outperforms fancy architectures through better negative sampling
 - In-batch negatives, hard negatives, distillation more important than model size

ZERO-SHOT PERFORMANCE IS CRITICAL

- Generalization to new domains as important as in-domain accuracy
- Sparse models (SPLADE++) often generalize better than dense models
- Bottleneck pre-training (SIMLM) improves transfer

SECURITY CANNOT BE IGNORED

- 0.02% adversarial content = 60% attack success rate
- Production systems need active defenses

CHALLENGES AND FUTURE DIRECTIONS



CHALLENGES



01

SECURITY AND ROBUSTNESS

02

MULTIMODAL CAPABILITIES

03

BIAS AND FAIRNESS

04

SUSTAINABILITY

05

EVALUATION BEYOND BENCHMARKS

FUTURE WORK

01

DEVELOP ROBUST DEFENSES AGAINST
ADVERSARIAL ATTACKS

02

MULTIMODAL RAG INTEGRATING TEXT,
IMAGES, AND CODE

03

OPTIMIZE HARDWARE ACCELERATION
FOR RAG SYSTEMS



CONCLUSION





CONCLUSION

Semantic search has evolved from keyword-based matching to dense vector retrieval and now to tightly integrated retrieval-augmented generation systems. Early advances focused on representation learning and contrastive training. Subsequent research addressed scalability through compression, indexing, and hardware acceleration. More recent work integrates retrieval directly into LLM pipelines, enabling adaptive, grounded generation. However, challenges remain in robustness, security, and billion-scale deployment. The boundary between retrieval and generation is increasingly dissolving, suggesting that future systems will not treat them as separate stages, but as unified, adaptive semantic reasoning frameworks.



MBA519: BIG DATA ANALYTICS



THANK YOU ANY QUESTIONS?

2025-2026