MBA519: BIG DATA ANALYTICS

TUNIS BUSINESS SCHOOL

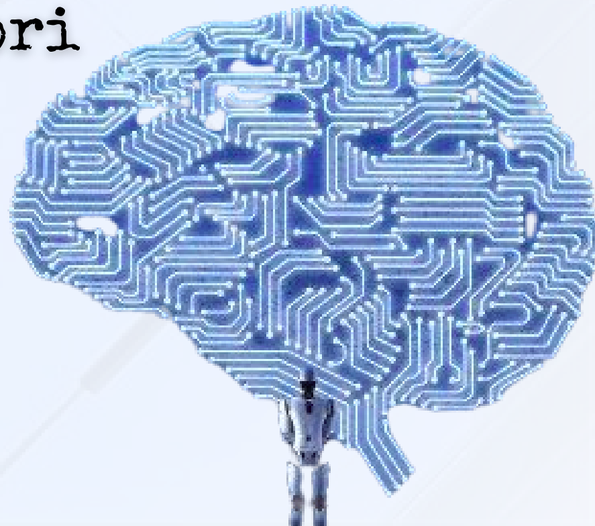
UNIVERSITÉ DE TUNIS

SEMANTIC SEARCH ENGINES OVER BIG DATA USING LLMS

SEARCHING LARGE TEXT CORPORA WITH CONTEXTUAL UNDERSTANDING

Khouloud Ben Younes

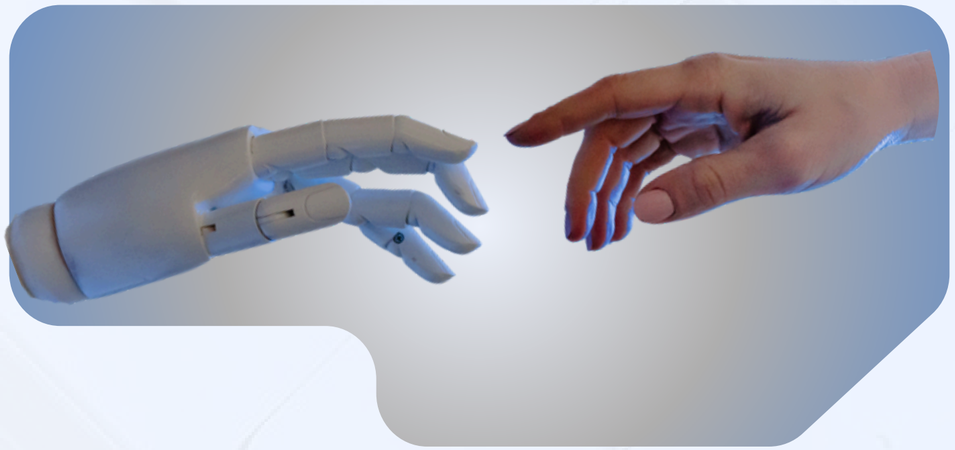
Montaha Ghabri



Evaluated by :
Dr. Manel AbdelKader

2025-2026

AGENDA



01 INTRODUCTION

02 LITERATURE SELECTION

03 BACKGROUND & KEY CONCEPTS

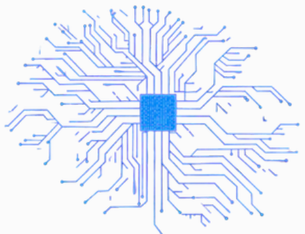
04 REVIEW & ANALYSIS OF EXISTING WORK

05 MAJOR INSIGHTS AND DISCOVERIES

06 CHALLENGES AND FUTURE DIRECTION

07 CONCLUSION

MOTIVATION



Start

1 MOTIVATION

Problem

File Edit Search Help

- Explosive growth of digital text data
- Keyword search (BM25) suffers from vocabulary mismatch
- LLMs hallucinate without grounding

Importance & Relevance

File Edit Search Help

- Semantic understanding at scale
- Efficient billion-scale search
- Reliable integration with LLM generation

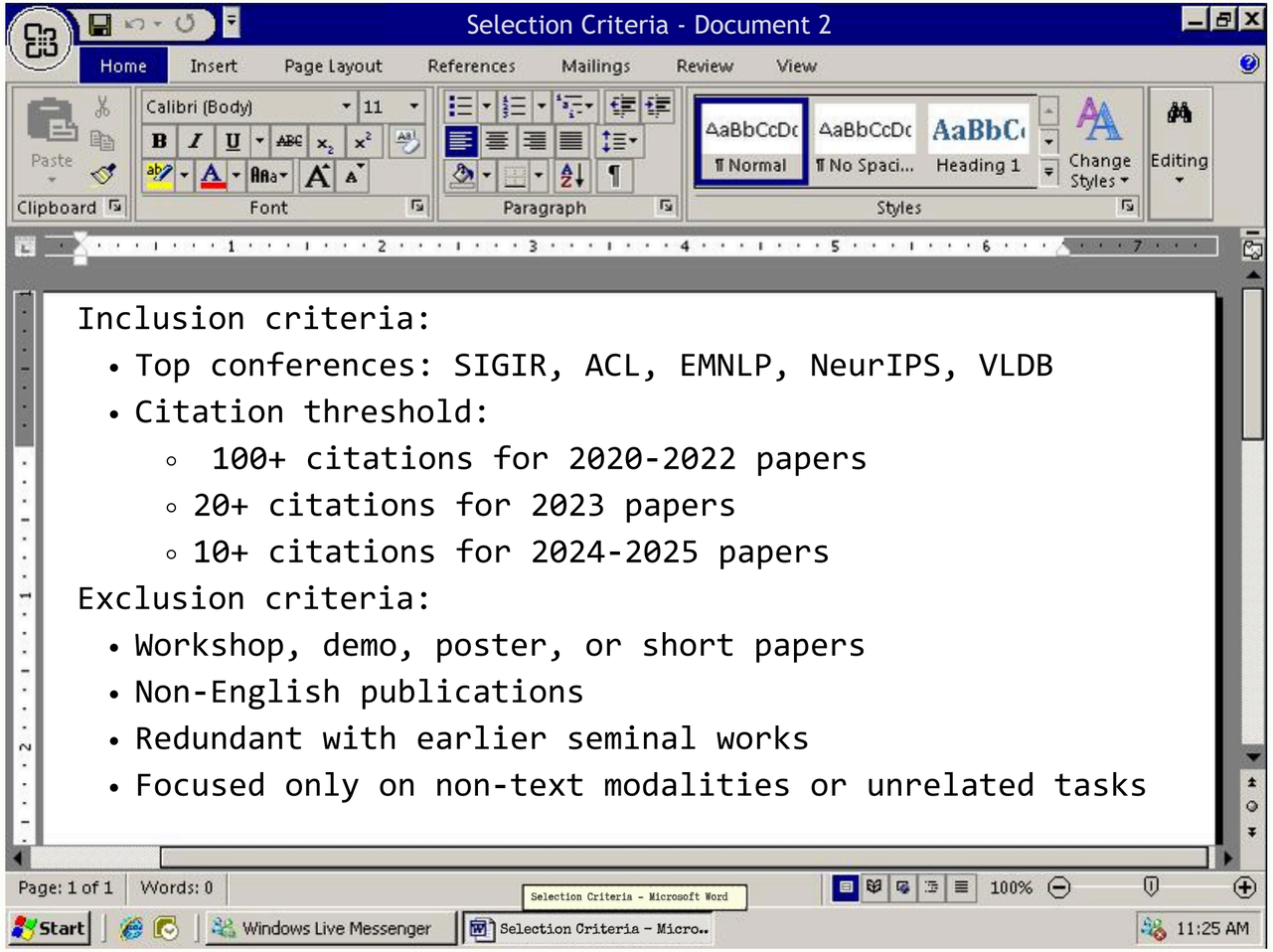
Research Questions

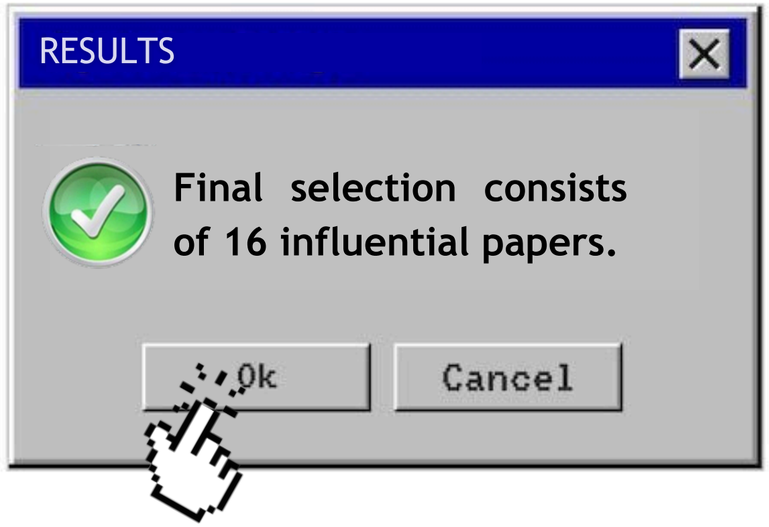
File Edit Search Help

- RQ1: compares sparse, dense, and hybrid paradigms.
- RQ2: evaluates training methodologies like negative sampling.
- RQ3: addresses billion-scale deployment challenges.
- RQ4: examines security vulnerabilities like poisoning.
- RQ5: explores RAG integration with hardware.

LITERATURE SELECTION

2 LITERATURE SELECTION

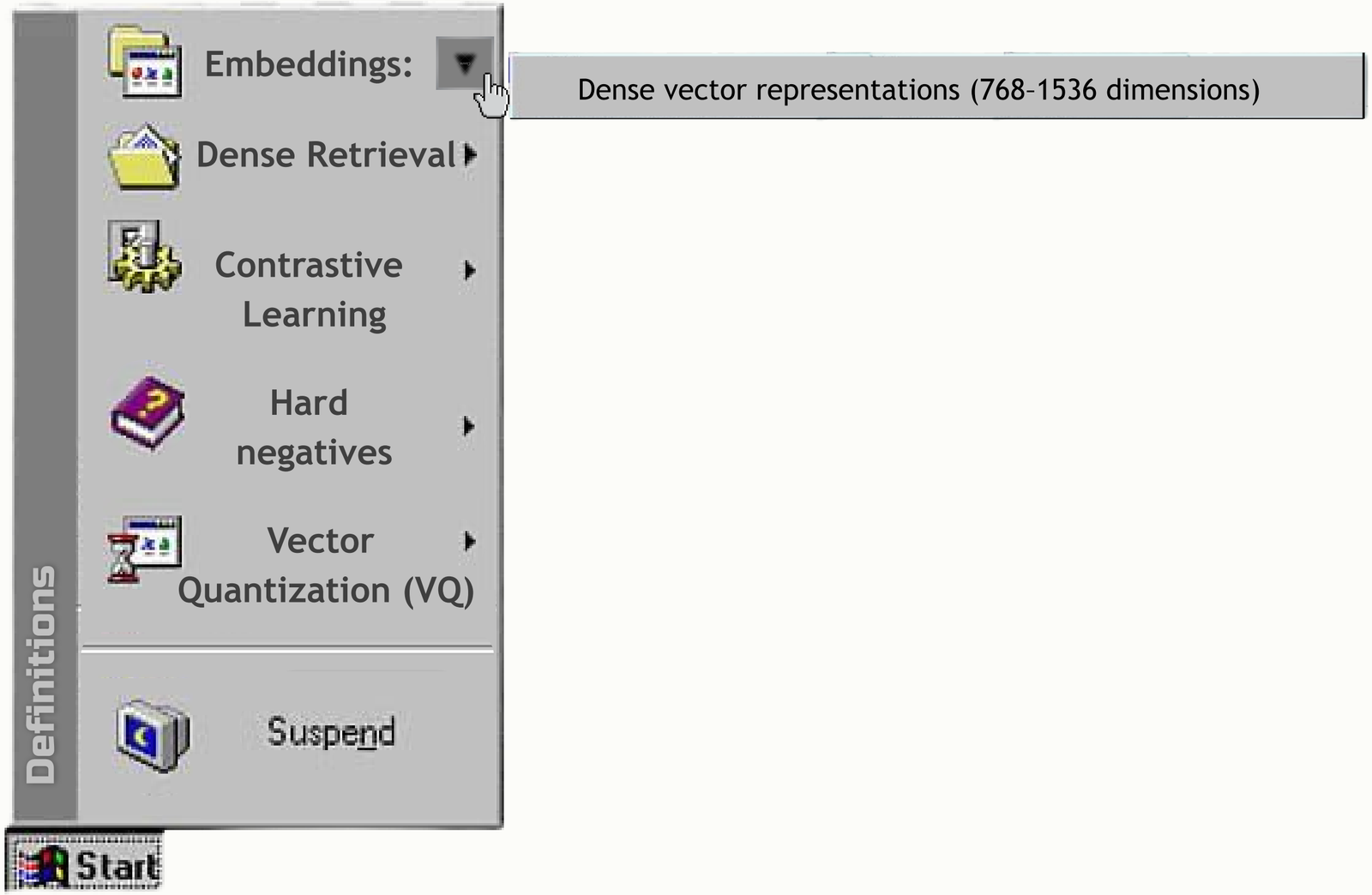




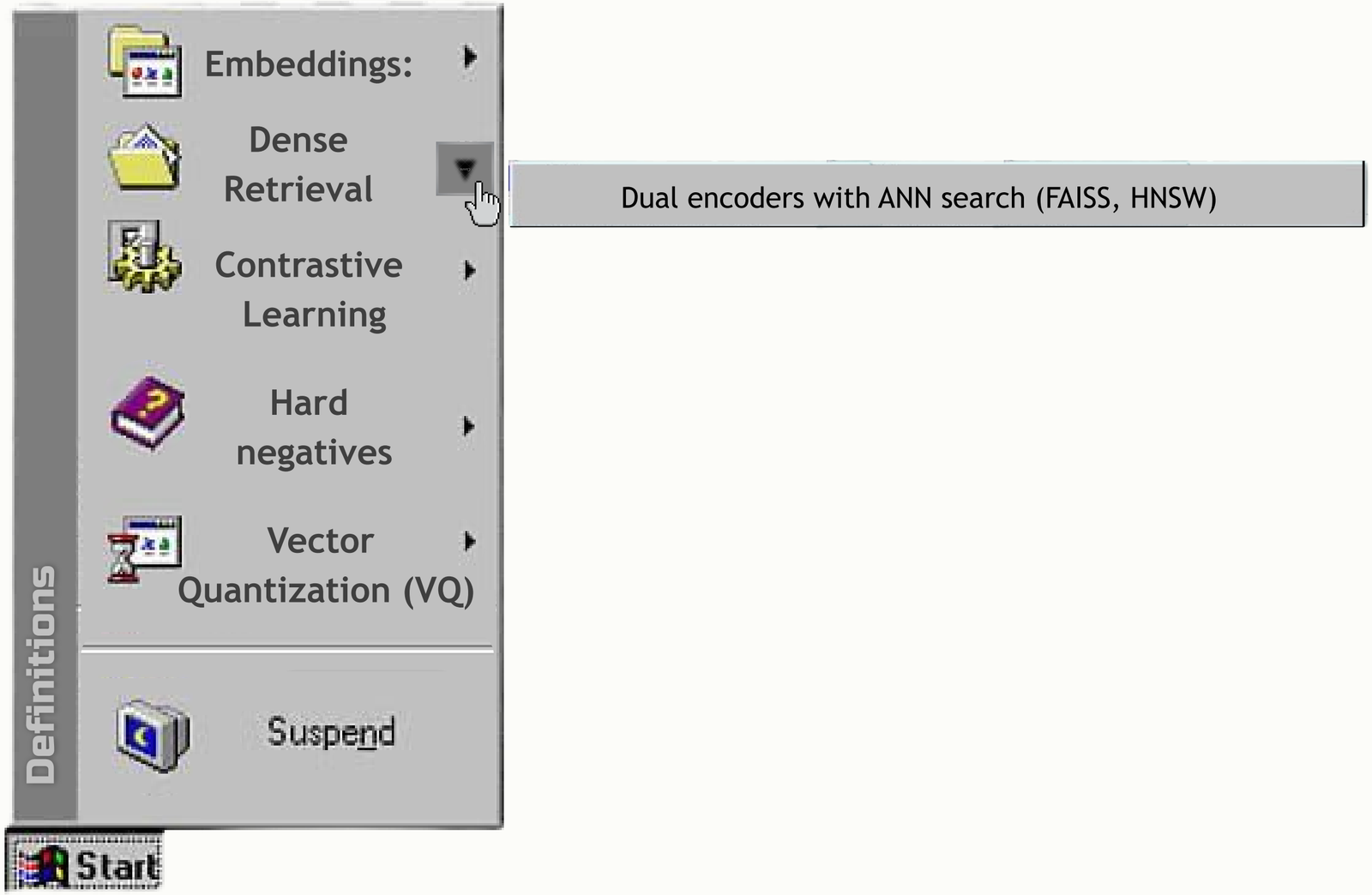
Title	Authors	Year	Citations	Category	Link	Database	Status	Priority	Notes
Dense Passage Retrieval for Open-Domain Question Answering	Vladimir Karpukhin• , Barlas Oğ	2020	5695	Dense Retrieval Foundations	https://arxiv.org/pdf/2004.0496				
SIMLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval	Liang Wang, Nan Yang, Xiaolo	2023	148	Dense Retrieval Foundations	https://aclanthology.org/20				

BACKGROUND & KEY CONCEPTS

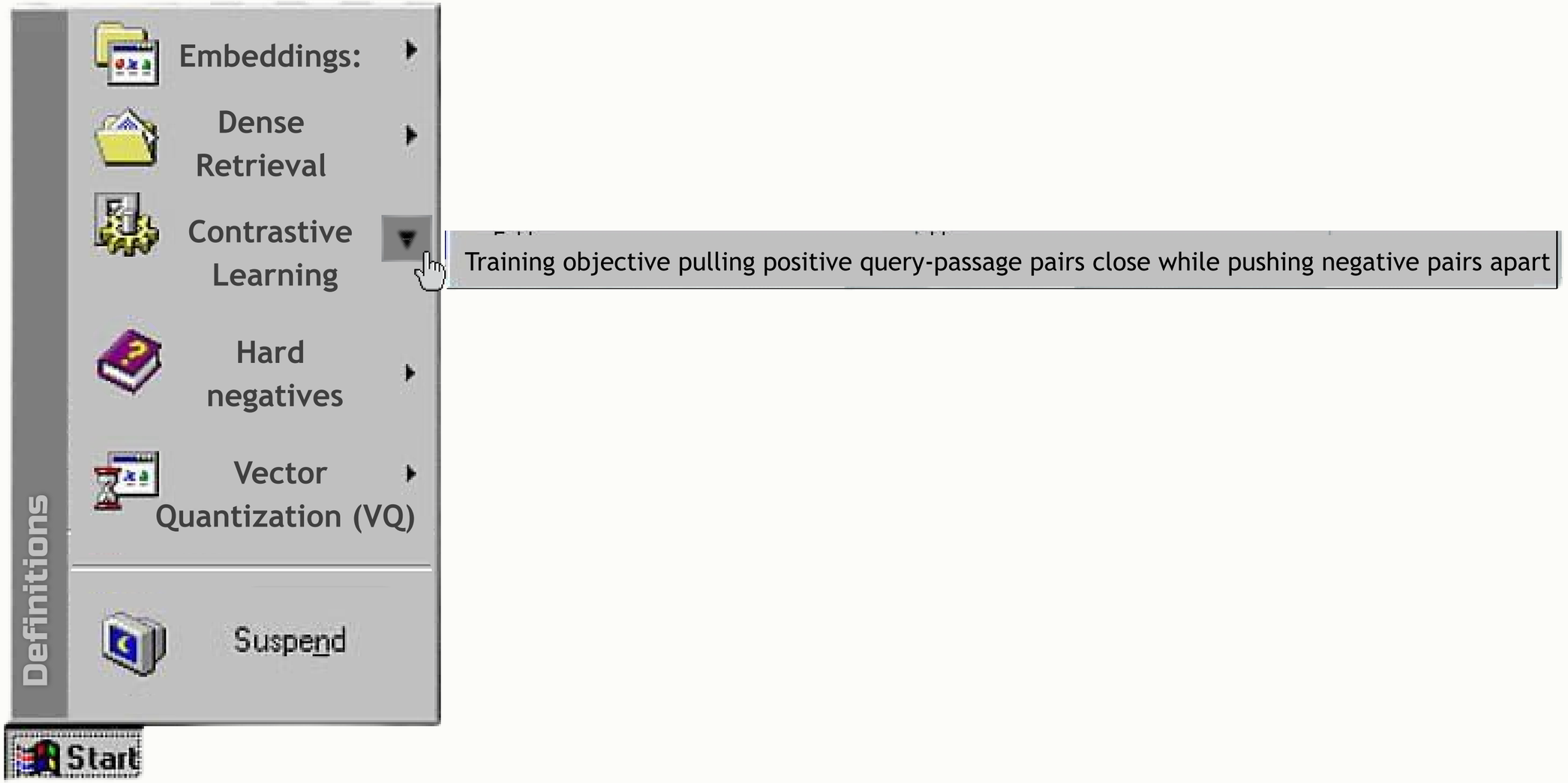
3 Background & Key Concepts



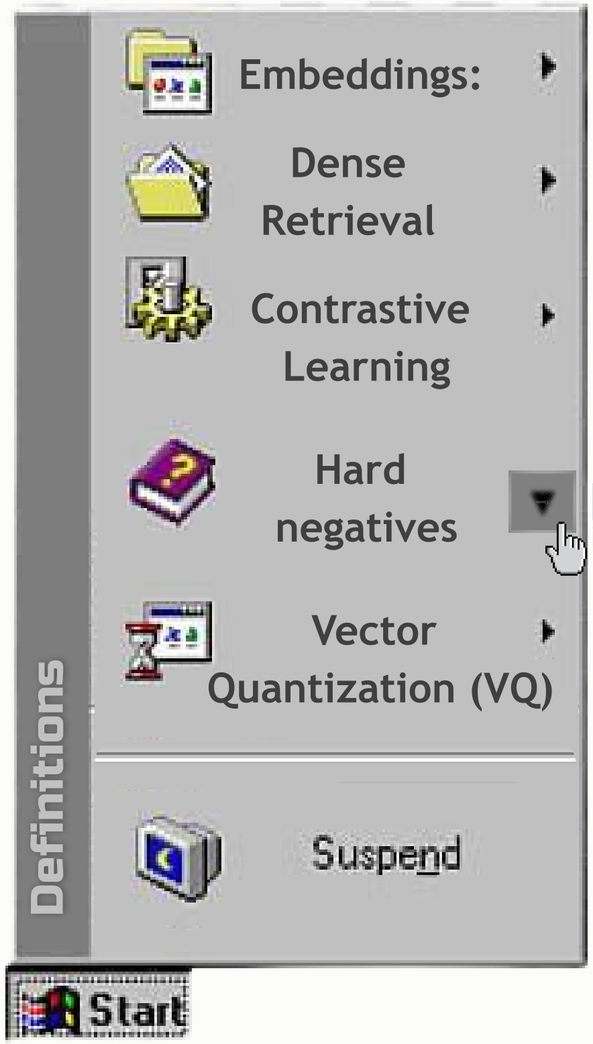
3 Background & Key Concepts



3 Background & Key Concepts

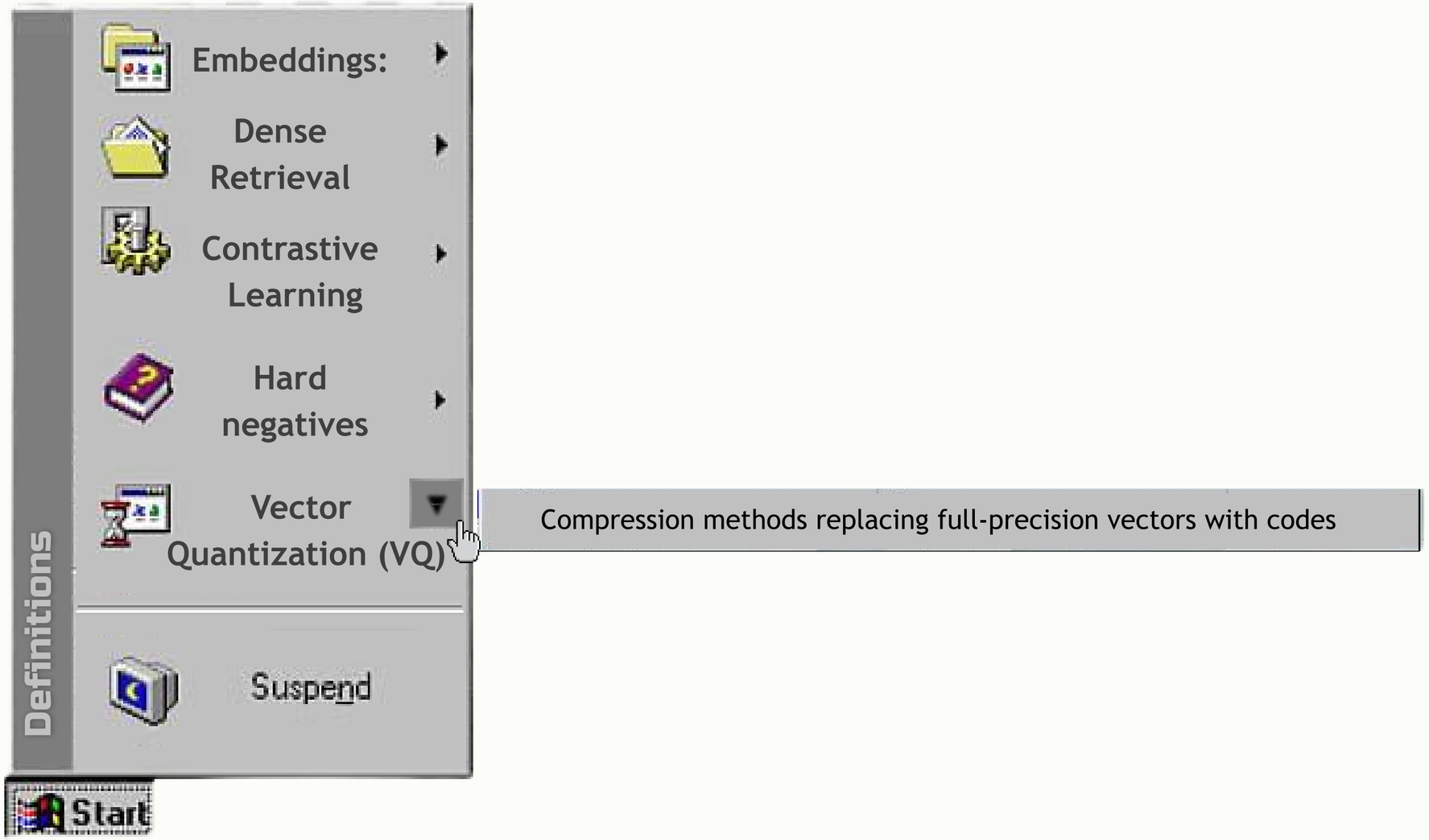


3 Background & Key Concepts

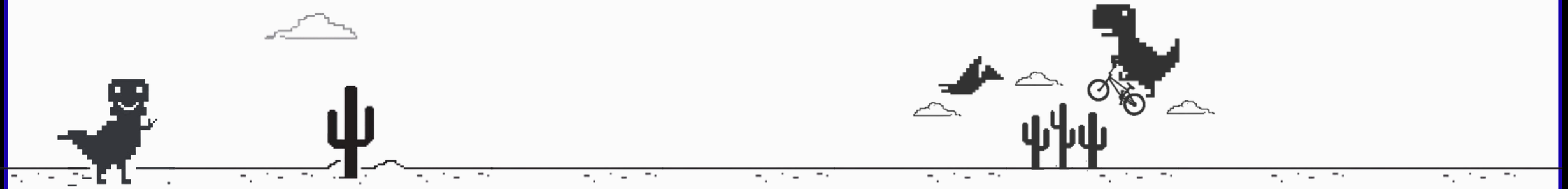


semantically similar but incorrect passages that improve discrimination

3 Background & Key Concepts



REVIEW OF EXISTING WORK



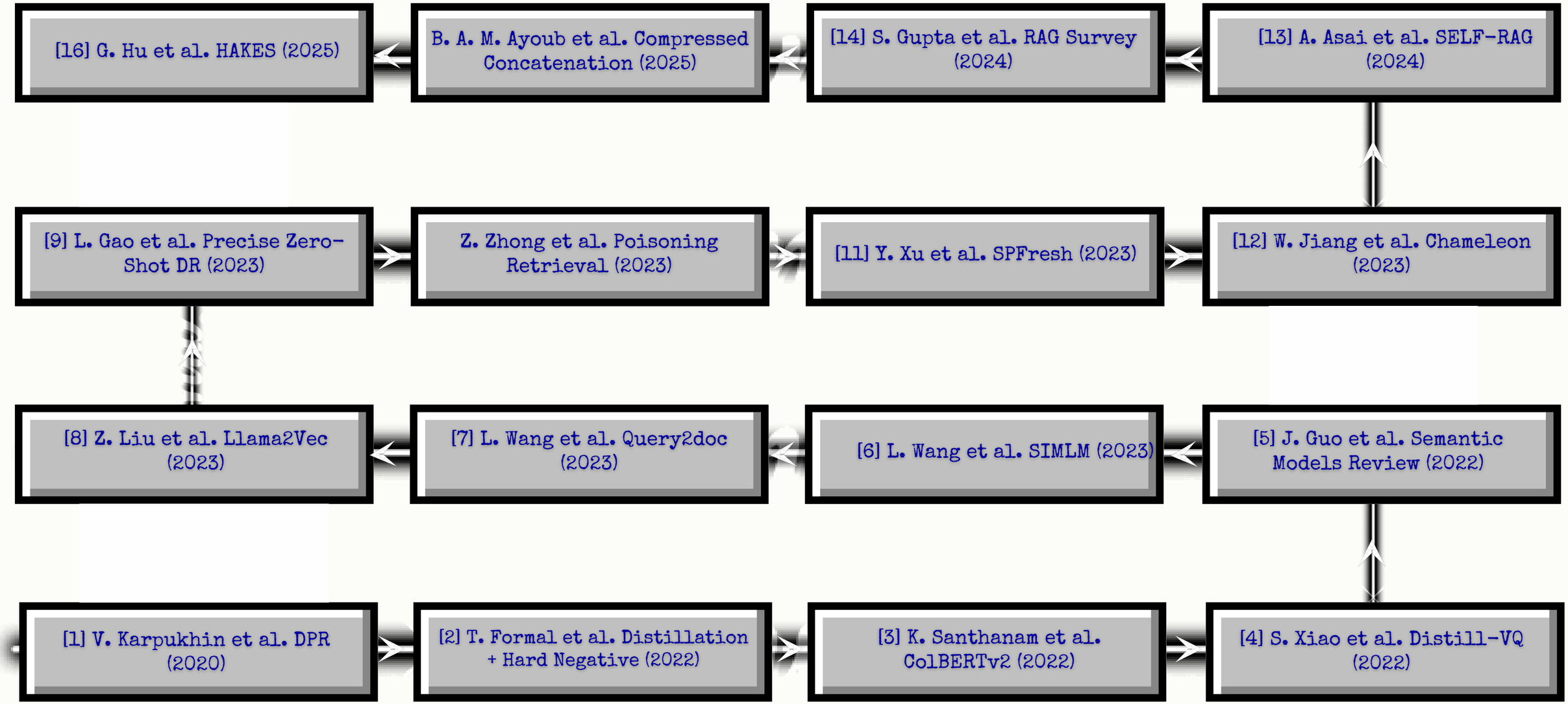
You are offmark

Try:

Reading recent papers.

FOUR RESEARCH CATEGORIES

TIMELINE


[1] V. Karpukhin et al. DPR (2020)
[2] T. Formal et al. Distillation + Hard Negative (2022)
[3] K. Santhanam et al. ColBERTv2 (2022)
[4] S. Xiao et al. Distill-VQ (2022)
[5] J. Guo et al. Semantic Models Review (2022)
[6] L. Wang et al. SIMLM (2023)
[7] L. Wang et al. Query2doc (2023)
[8] Z. Liu et al. Llama2Vec (2023)
[9] L. Gao et al. Precise Zero-Shot DR (2023)
Z. Zhong et al. Poisoning Retrieval (2023)
[11] Y. Xu et al. SPFresh (2023)
[12] W. Jiang et al. Chameleon (2023)
[13] A. Asai et al. SELF-RAG (2024)
[14] S. Gupta et al. RAG Survey (2024)
B. A. M. Ayoub et al. Compressed Concatenation (2025)
[16] G. Hu et al. HAKES (2025)

1- DENSE RETRIEVAL FOUNDATIONS

DENSE PASSAGE RETRIEVAL (DPR, 2020)

File Edit Search Help

- Uses two separate BERT encoders (dual-encoder architecture, one for Query and one for passage
- Achieves 41% accuracy on Natural Questions dataset
- Uses in-batch negatives: Other questions' answers serve as hard negative
- 19+ percentage point improvement over traditional methods (BM25)

Vladimir Karpukhin et al., 2020

HOW NEURAL RETRIEVAL WORKS

Corpus poisoning attacks (SIMLM, 2023)

File Edit Search Help

- Uses bottleneck pre-training to learn document-level semantics

Pre-training is self-supervised (no labels-needed)

- Achieves strong zero-shot transfer across different domains

Zexuan Zhong, Ziqing Huang, Alexander Wettig, Danqi Chen, 2023

BETTER GENERALIZATION

Corpus poisoning attacks (SIMLM, 2023)

File Edit Search Help

- Adversary injects malicious passages into corpus
- Uses gradient-based optimization to craft adversarial text
- Attacks transfer across domains: 94% success on finance, 97% on forums

DEFENSE MECHANISMS: Embedding norm clipping: Reduces attack from 99.4% to 0.6%

Liang Wang et al, 2023

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

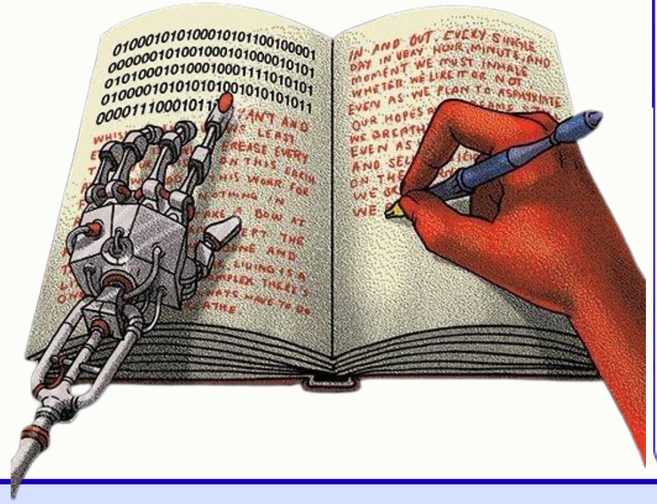
Yuming Xu et al., 2023

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

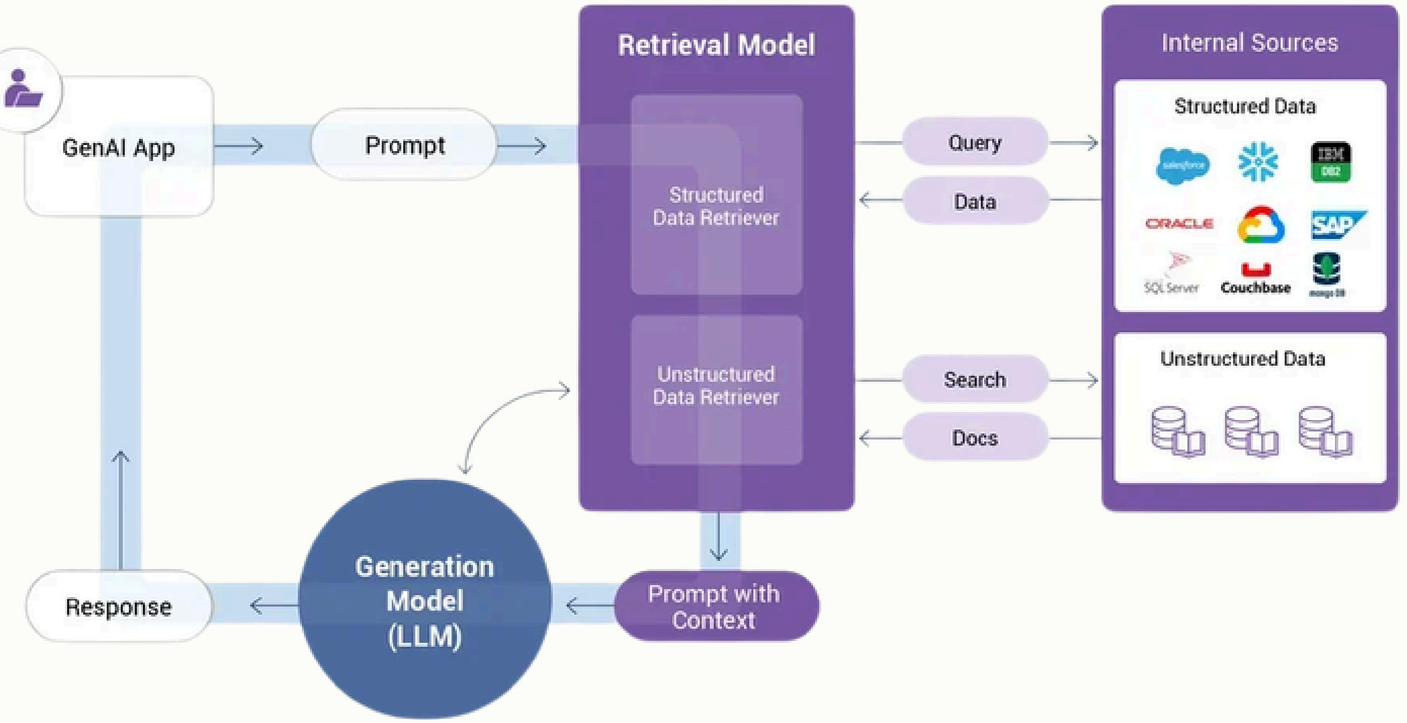
WHY RAG?

- 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

OK

COMBINE RETRIEVAL WITH LLM GENERATION TO REDUCE HALLUCINATIONS



3- RAG ARCHITECTURES

Combine retrieval with LLM generation to reduce hallucinations

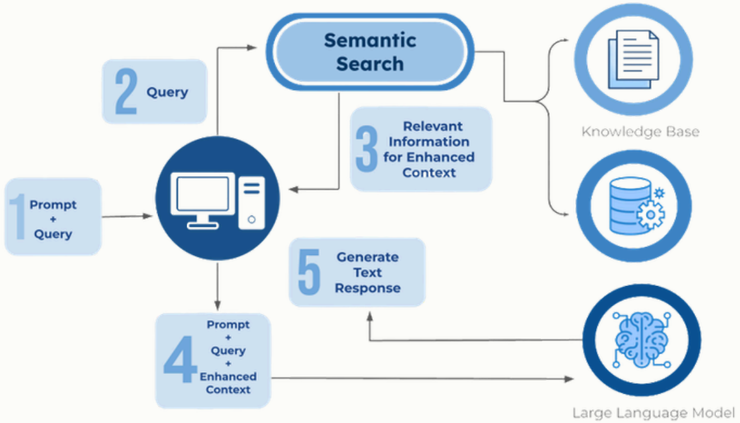
SELF-RAG - INTELLIGENT RETRIEVAL

- 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%

OK

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)

OK

Liang Wang, Nan Yang, Furu Wei, 2023

3- RAG ARCHITECTURES

Combine retrieval with LLM generation to reduce hallucinations

LLAMA2VEC

File Edit Search Help

- LLMs pre-trained for generation, not retrieval

STEP 1 - SYNTHETIC DATA GENERATION:

• For each passage in corpus, use Llama-2 to generate questions

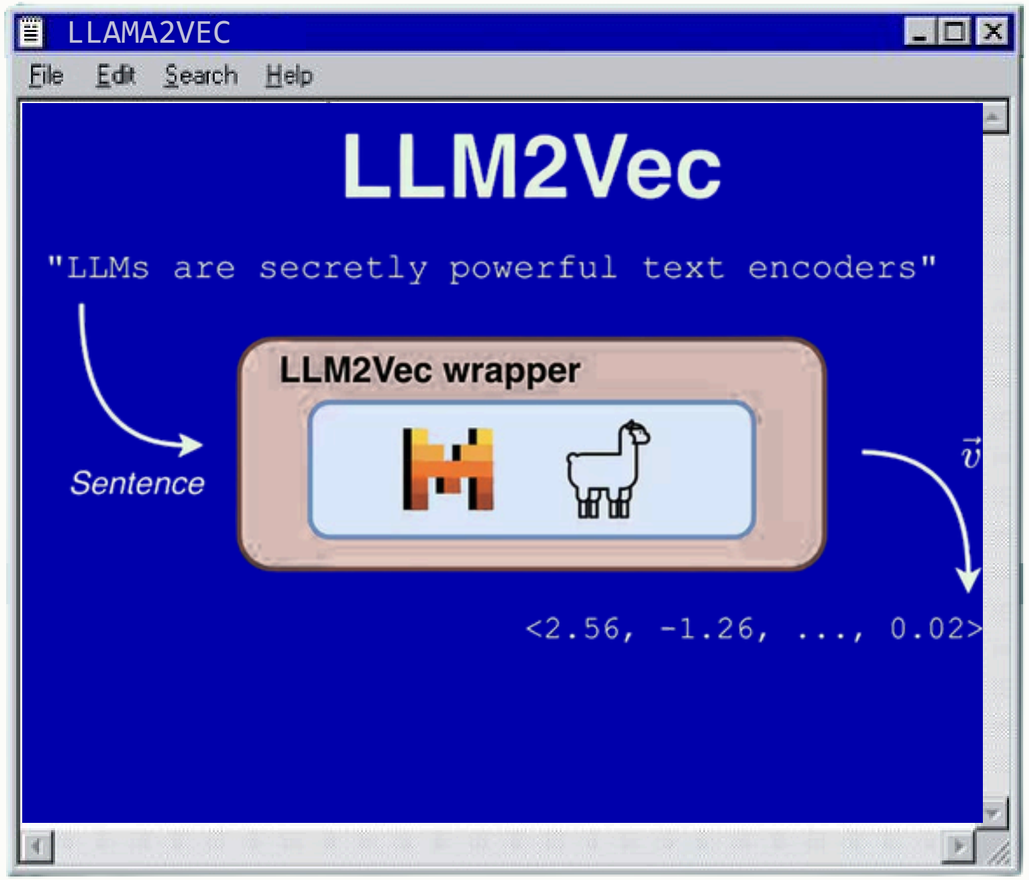
• Creates pseudo query-passage pairs without human annotation • Quality varies but provides useful training signal in aggregate

STEP 2 - CONTRASTIVE TRAINING:

• Train with standard contrastive loss on synthetic pairs • Maximize similarity for synthetic query + source passage • Minimize similarity to random passages from corpus

RESULTS:

Zheng Liu, Chaofan Li, Shitao Xiao, Yingxia Shao, Defu Lian, 2023



- 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

CHAMELEON - HARDWARE ACCELERATION:

- THE BOTTLENECK:

- **RAG has two distinct computational stages:**
  - **Retrieval: Vector search over large database**
  - **Generation: Transformer inference on LLM**

- **Different stages need different hardware optimizations.**

1. RETRIEVAL STAGE - FPGA ACCELERATION:

- FPGAs implement IVF-PQ with near-memory processing
- Memory organized to maximize bandwidth utilization
- Filters billions of vectors to find top candidates

2. GENERATION STAGE

- GPU ACCELERATION: GPUs handle transformer LLM inference

→ Processes retrieved documents + query through LLM

CPU COORDINATION: Orchestrates pipeline between FPGA and GPU:

- Receives queries → dispatches to FPGA
- Collects retrieved docs → batches for GPU

Wenqi Jiang, Marco Zeller, Roger Waleffe, Torsten Hoefler, Gustavo Alonso,2023

- 2.I6x latency reduction for end-to-end QA
- 3.I8x throughput improvement for concurrent requests

4- HYBRID & ADVANCED METHODS

SPLADE++ DENSE TRAINING FOR SPARSE MODELS

- 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**

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, Stéphane Clinchant

COLBERTV2

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

Keshav Santhanam, Omar Khattab, et al, 2022

HYDE - ZERO-SHOT HYPOTHETICAL DOCUMENTS

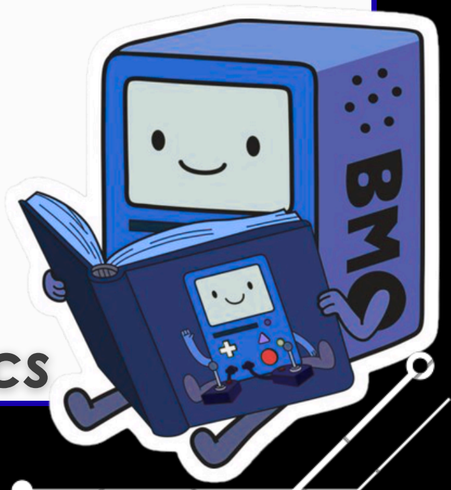
• Traditional:
Encode query → Search for similar documents
• HyDE: Generate hypothetical answer → Search for similar documents

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

Keshav Santhanam, Omar Khattab, et al, 202 - Luyu Gao et al, 20232

MAJOR INSIGHTS AND DISCOVERIES

Recent Directions and Emerging Topics



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

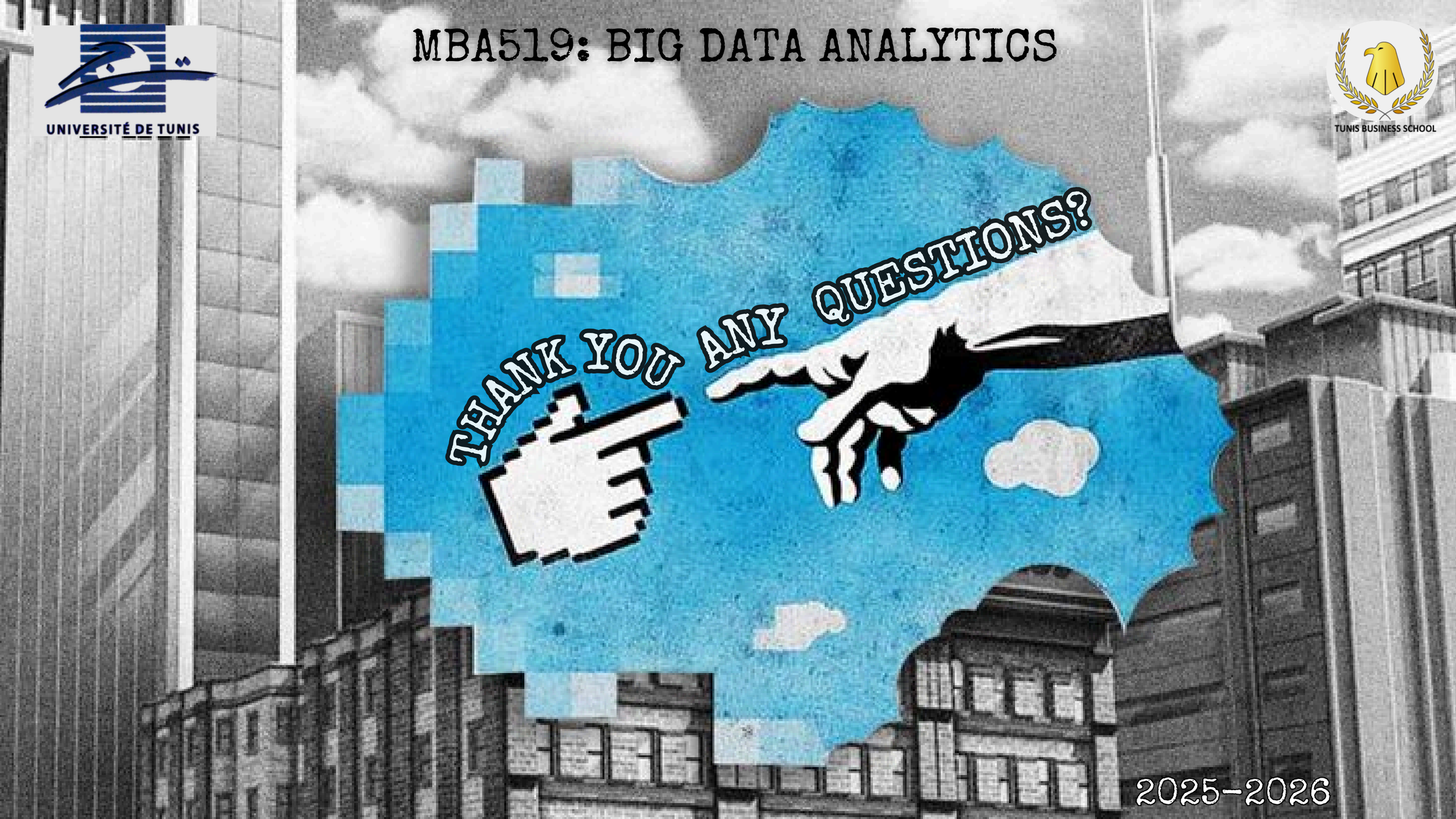
CONCLUSION

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.

UNIVERSITÉ DE TUNIS

TUNIS BUSINESS SCHOOL

MBA519: BIG DATA ANALYTICS



2025-2026