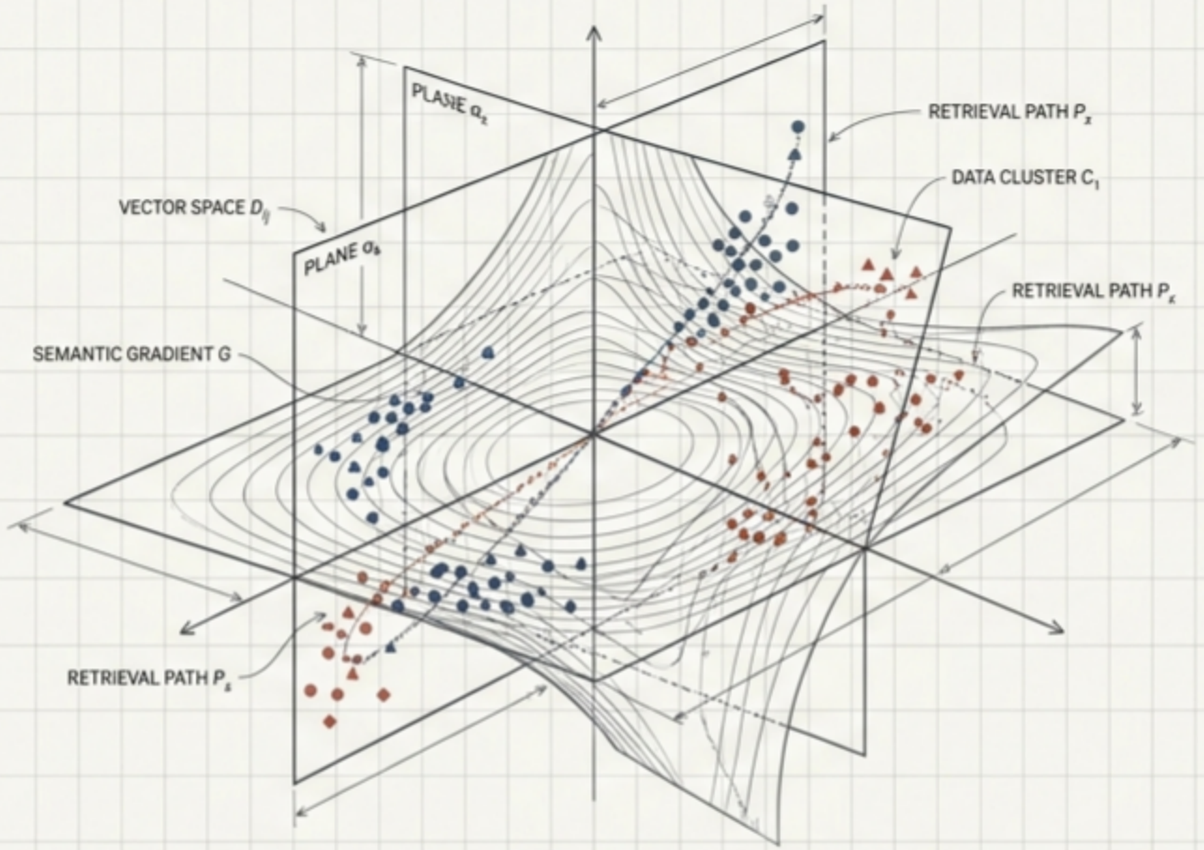


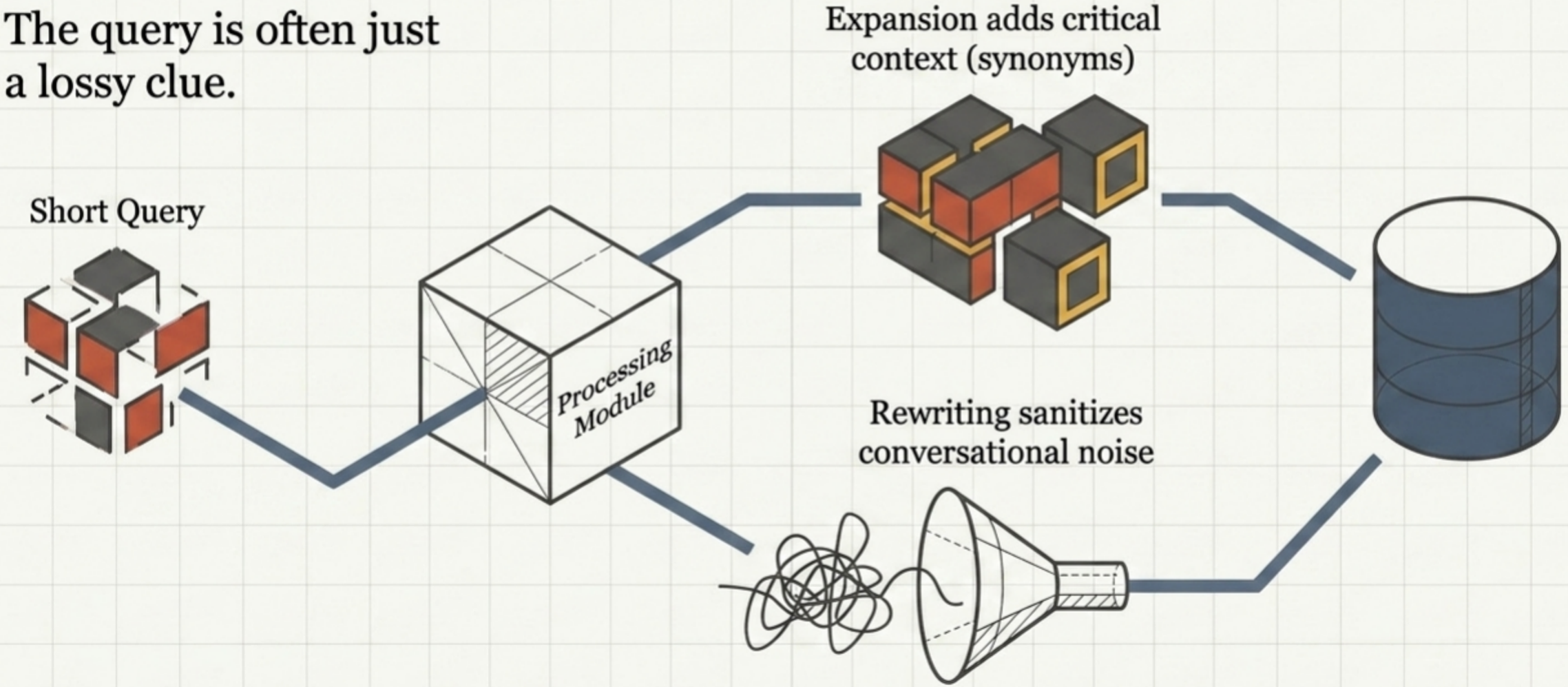
Pushing Semantic Search Beyond the Baseline

A curated dossier of advanced retrieval techniques—from query interventions to agentic loops—for large language models.

The Limits of the Standard Pipeline

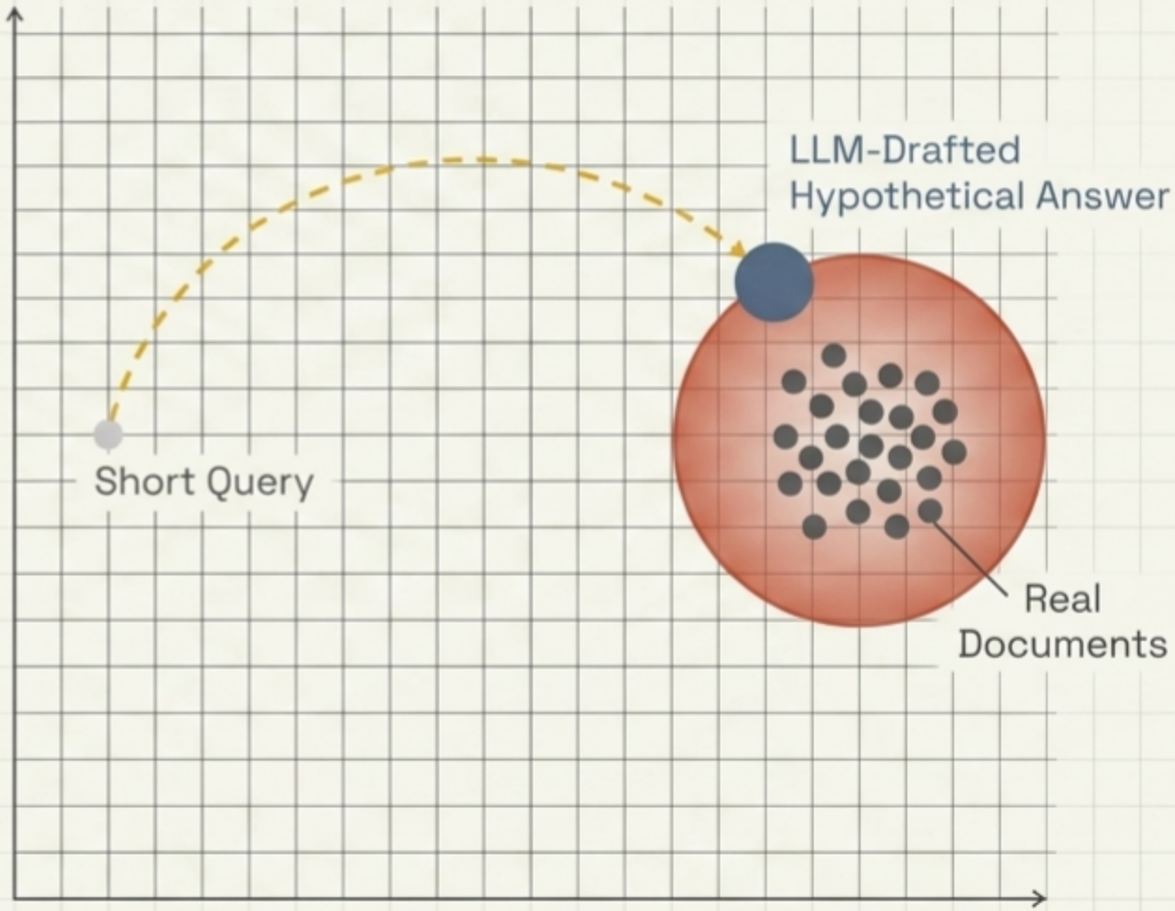
Intervening Before the Database

The query is often just a lossy clue.



Bridging the Semantic Gap with HyDE

- HyDE (Hypothetical Document Embeddings).
- Instead of searching with a short question, ask an LLM to draft a hypothetical answer.
- Retrieve using that draft. It sits fundamentally closer to the real answer passages in the vector space.



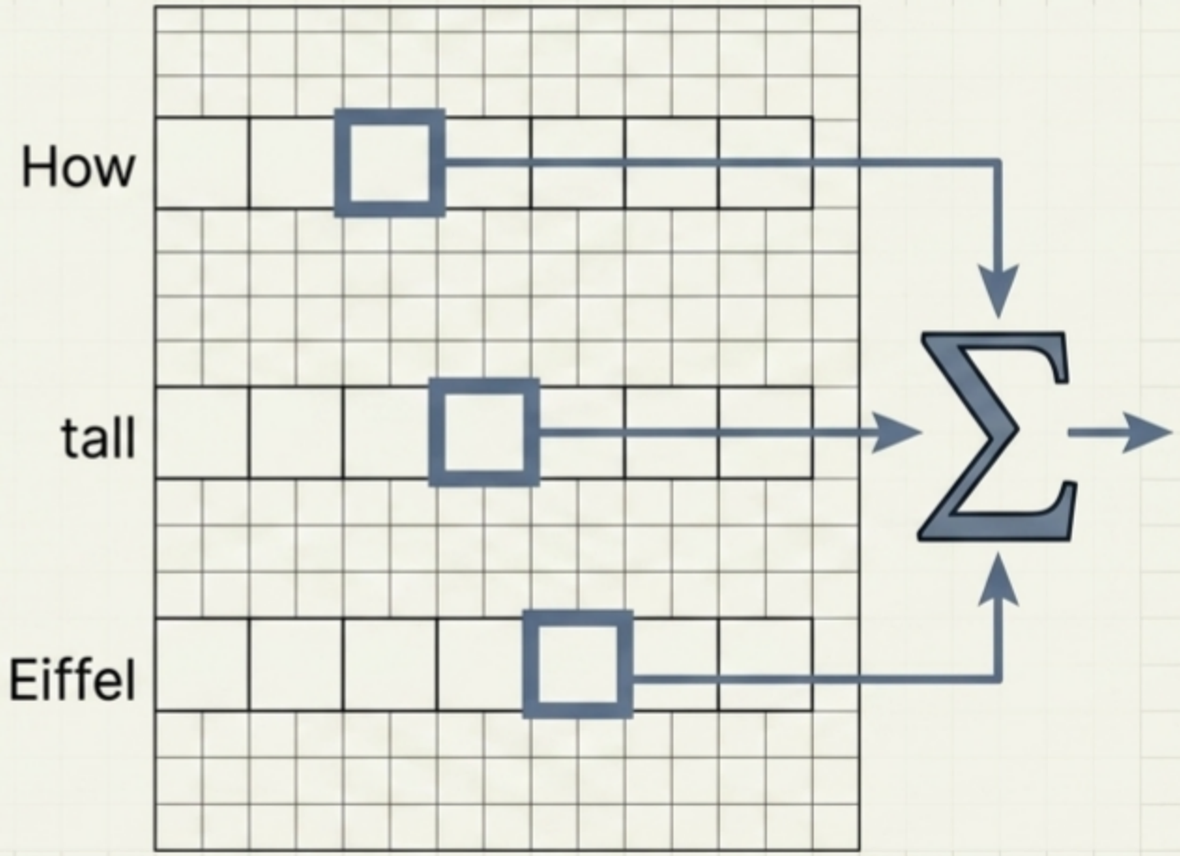
Rethinking the Architecture for Granularity

ColBERT bridges the gap. It keeps high-resolution token detail without sacrificing offline pre-computability.

	Bi-encoder	Cross-encoder	**ColBERT (Late Interaction)**
Embedding Granularity	Document	Token	**Token**
Pre-computability	Yes	No	**Yes**
Speed	Fast	Slow	**Fast**
Accuracy	Baseline	High	**High**

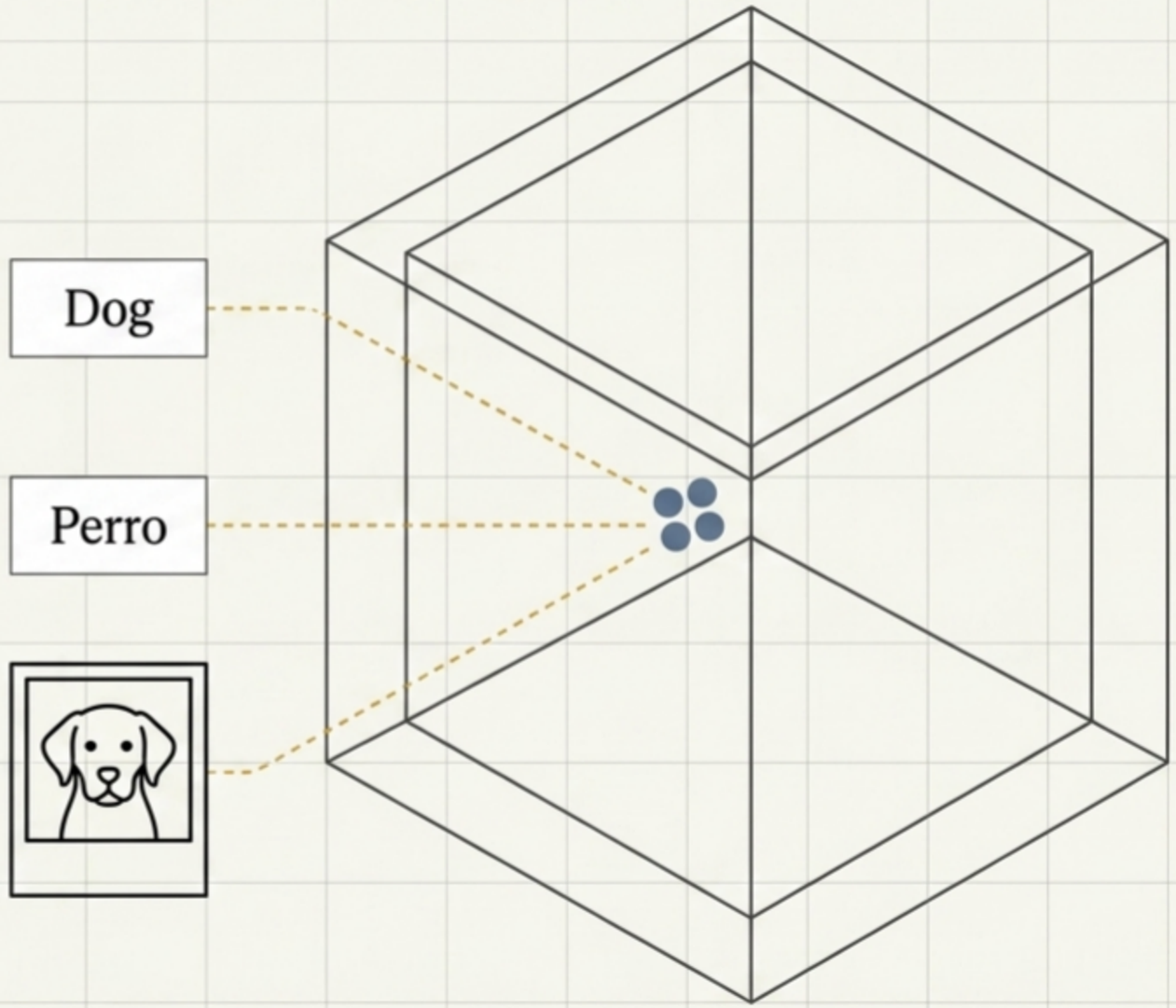
Late Interaction via MaxSim Scoring

- ColBERT represents text as one vector per token.
- MaxSim Mechanism: For each query token, find its best-matching document token.
- Sum these maximums together to get the final score.

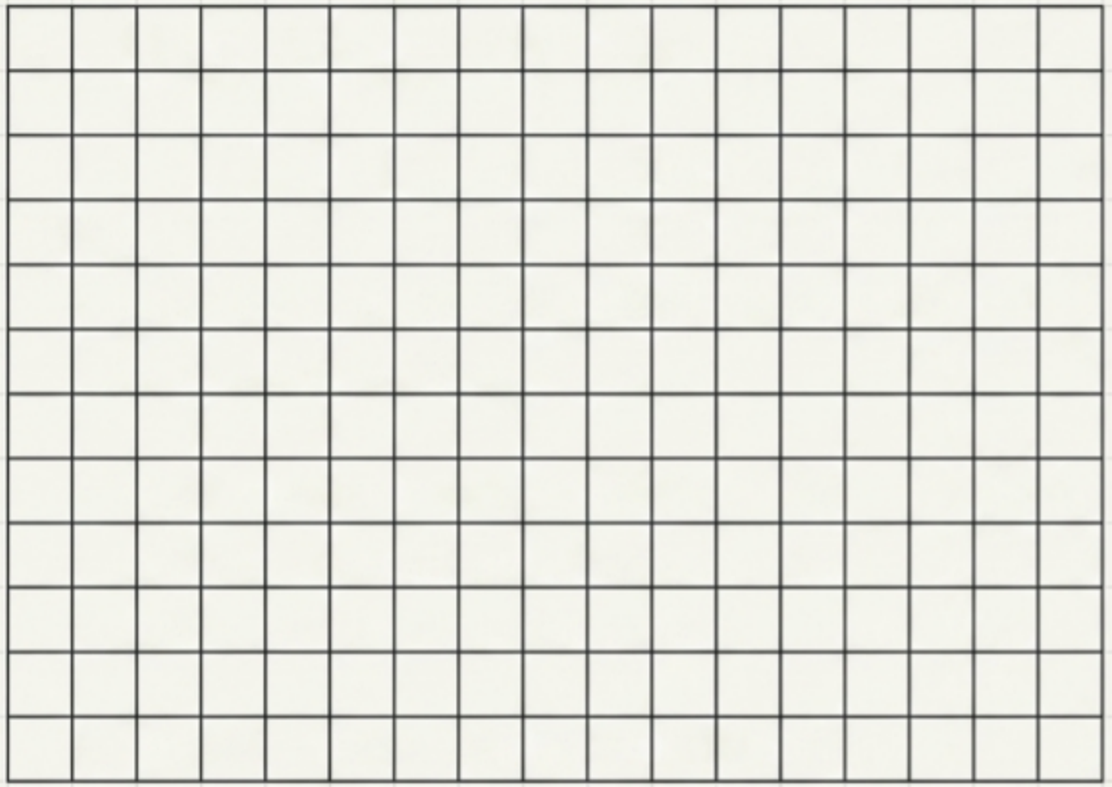


Expanding Modality in the Vector Space

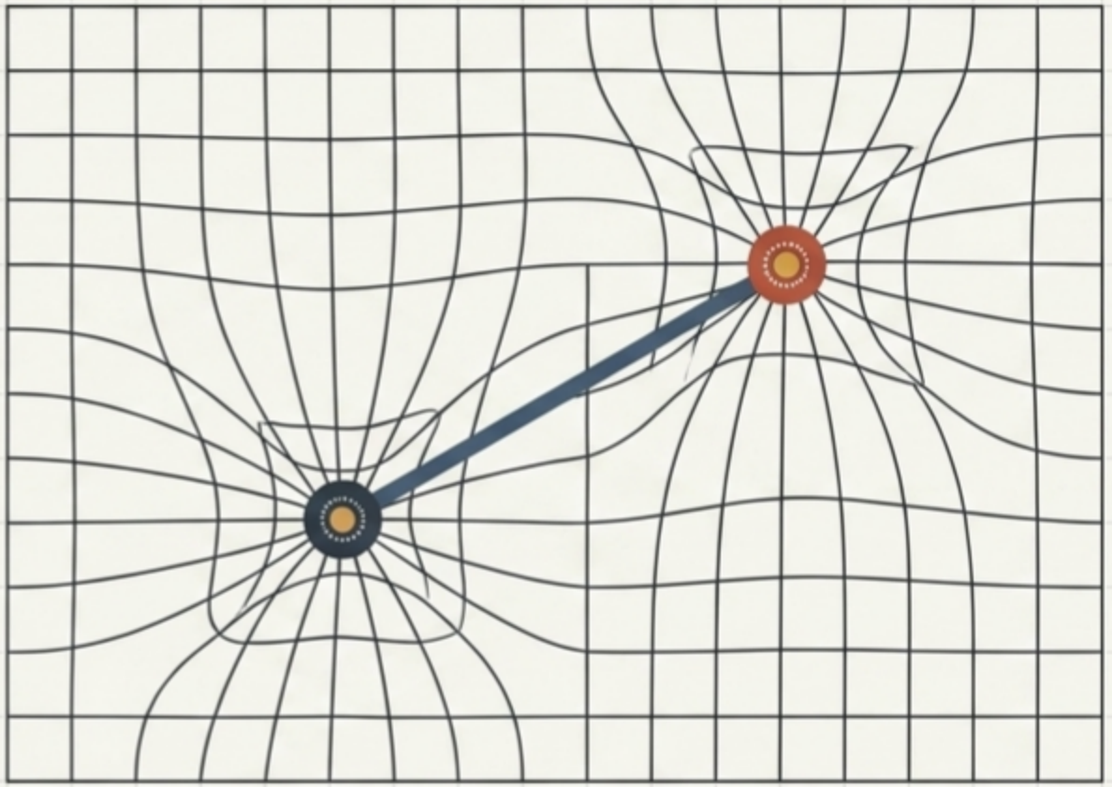
- Multilingual models align concepts regardless of language.
- Multimodal models (like CLIP) project images and text into the exact same semantic space.
- Result: A text query can directly retrieve images using the exact same nearest-neighbor machinery.



Shaping the Space to Your Domain



Pre-trained Space



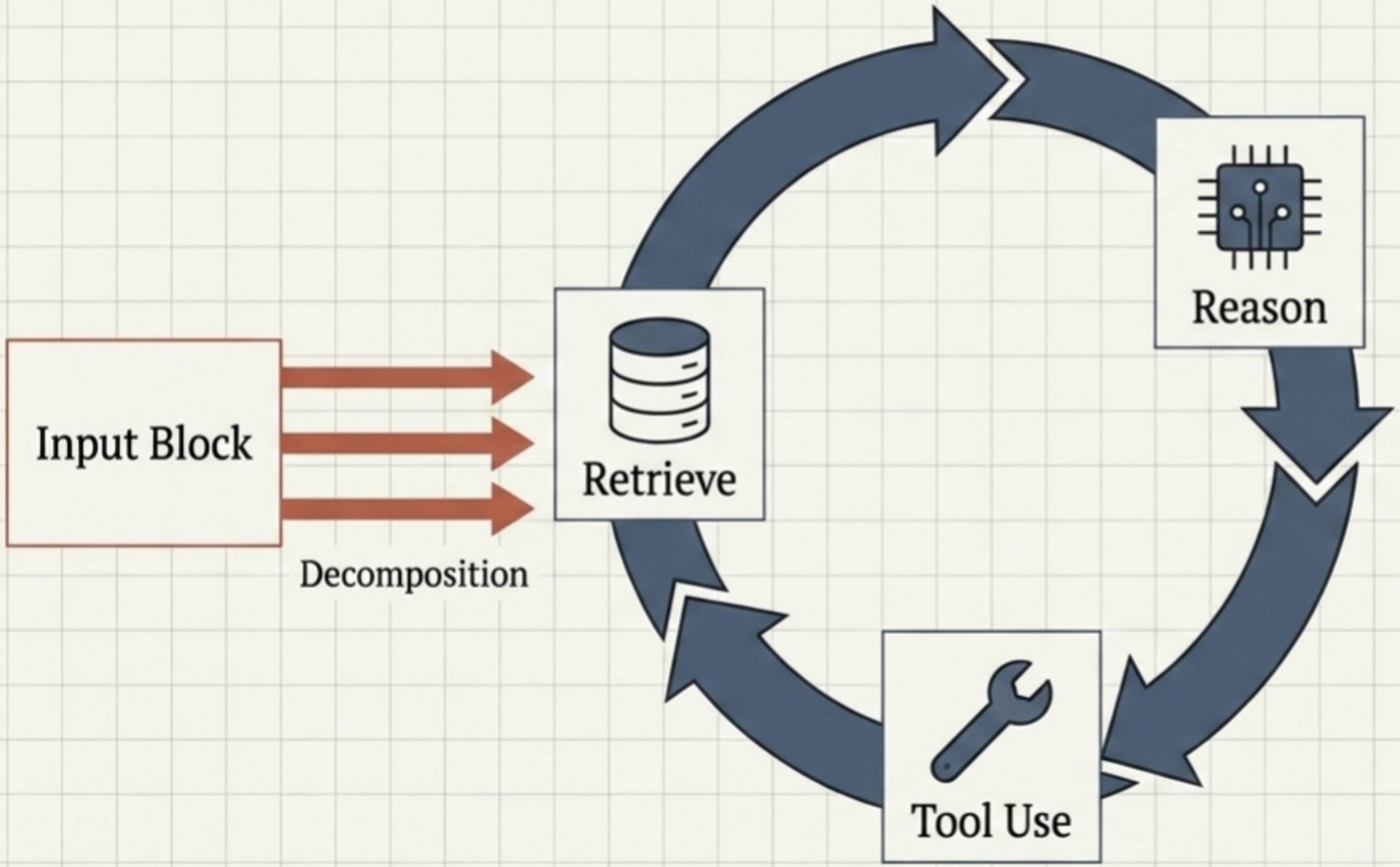
Fine-tuned Domain Space

Fine-tuning Embeddings: Training an embedding model on your specific domain data.

The model learns to physically reshape the semantic space, pulling domain-specific relevant passages closer to their corresponding queries.

Trade-off: Costs significant training data to execute, but yields substantial, permanent quality lifts.

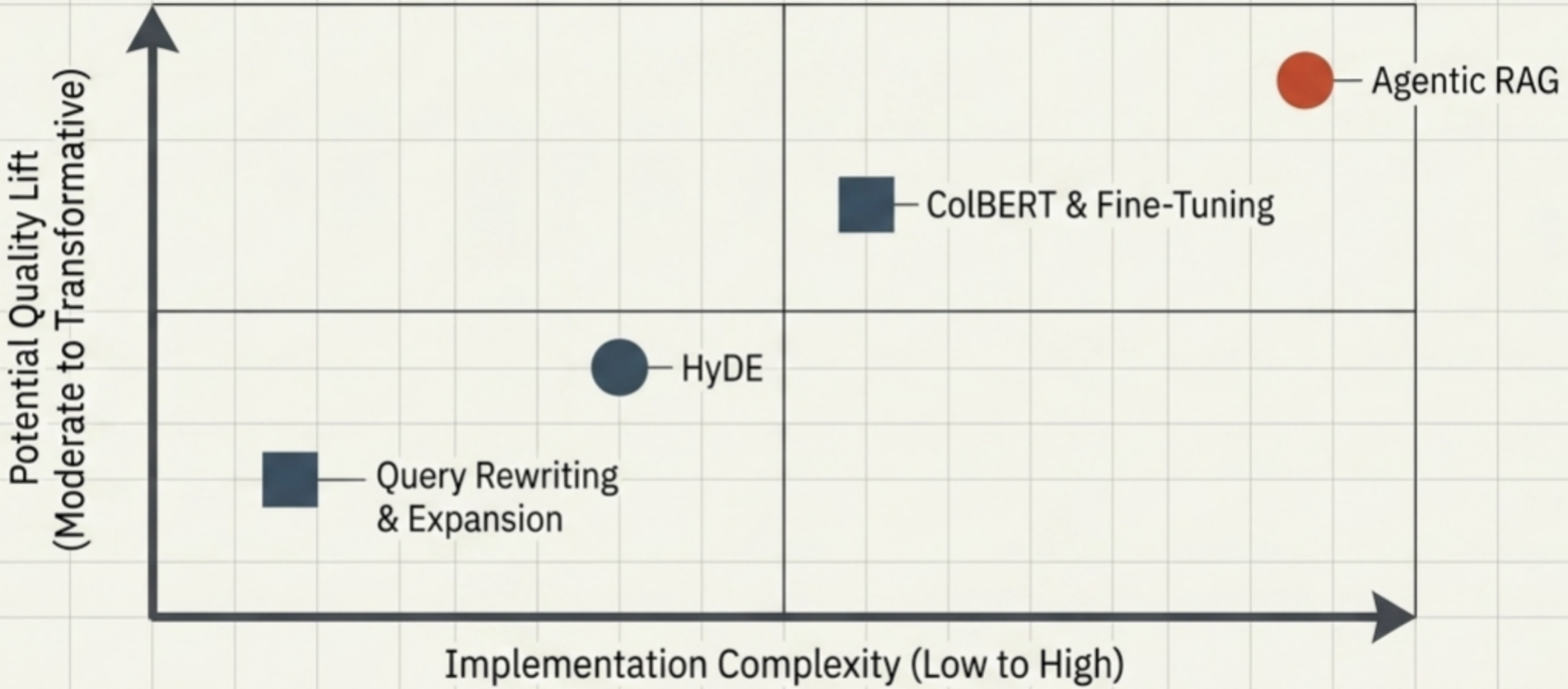
Iterative Loops for Complex Queries



Agentic / Multi-hop RAG is designed for hard questions.

- Breaks down complex queries into multiple lookups.
- Highly powerful, but introduces new trade-offs: harder to control, increased latency, and higher token costs.

The ROI Matrix: Measure Before Adopting



Every advanced technique adds system weight. None should be adopted blindly.
Rule of thumb: Each intervention must earn its place by demonstrably improving baseline retrieval metrics on YOUR specific data.

Knowledge Check

Explain HyDE in one sentence and why it help and why it helps.

Retrieve using an LLM-drafted hypothetical answer, which sits closer to real answer passages than a short question.

What does ColBERT keep that a bi-encoder throws away, and how does MaxSim MaxSim use it?

Per-token vectors; MaxSim sums each query token's best-matching document token.

Why can ColBERT precompute document vectors while a cross-encoder can't?

ColBERT embeds document tokens independently of the query; a cross-encoder must read them jointly.