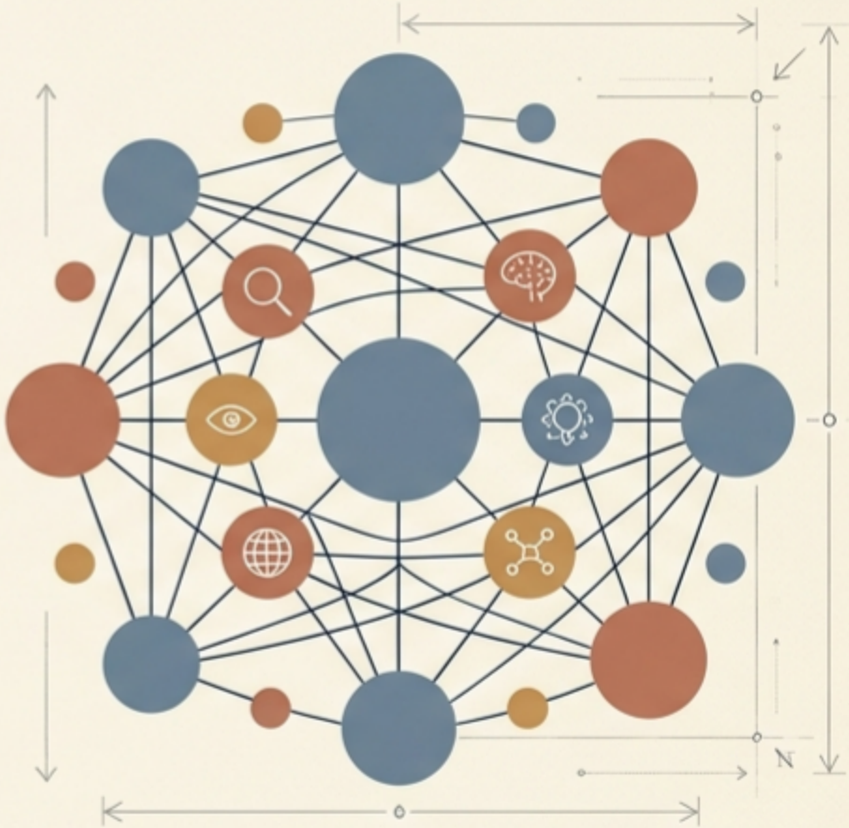
Search Semantically: Beyond Keywords

How Large Language Models transformed search from **matching literal words** to **understanding underlying meaning.**

LITERAL KEYWORD MATCHING

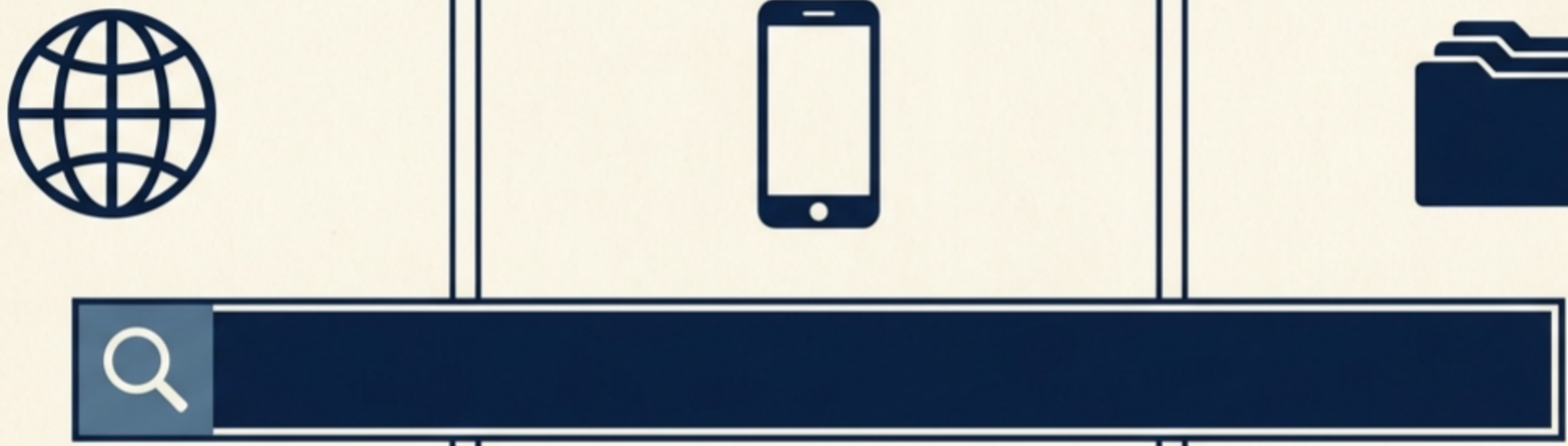


UNDERLYING MEANING (SEMANTIC SEARCH)

Understanding context, intent, and relationships between concepts.

The Universal Navigation Engine

Search is the fundamental way we navigate the digital world.



The Web
(Global Knowledge)

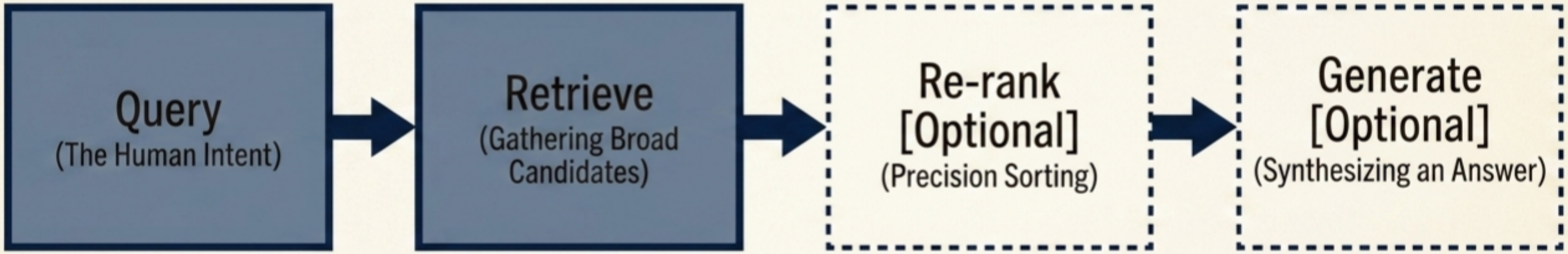
Mobile Apps
(Targeted Utilities)

Enterprise Documents
(Internal Context)

The Core Problem: Literal vs. Latent Meaning

	Keyword Search	Semantic Search
Matching Mechanism	Matches literal words and spelling.	Matches underlying meaning and intent.
Handling of Context	Blind to context; misses synonyms.	Connects related concepts effortlessly.
Query Example: “stop my laptop dying”	✗ Returns articles about “dying clothes” (Fails).	✓ Returns guides on “battery life” (Succeeds).

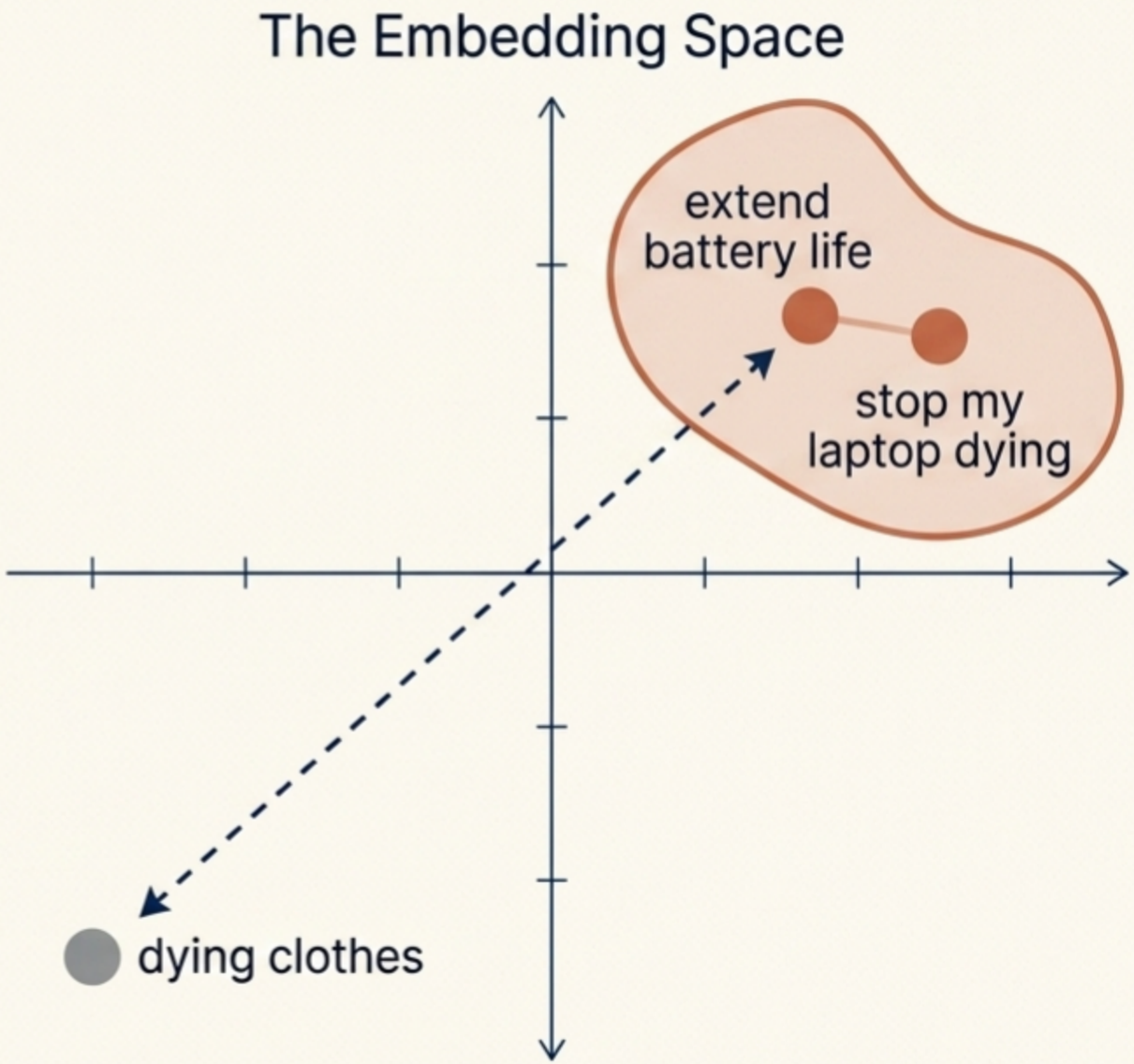
The 4-Step Semantic Pipeline



Step 1 & 2: Query & Retrieve (Mapping Meaning)

Instead of looking for shared letters, semantic retrieval calculates the mathematical distance between concepts.

Key Insight: We quickly pull a handful of likely-relevant documents from a massive corpus using embeddings—transforming text into coordinates so machines can match meaning.

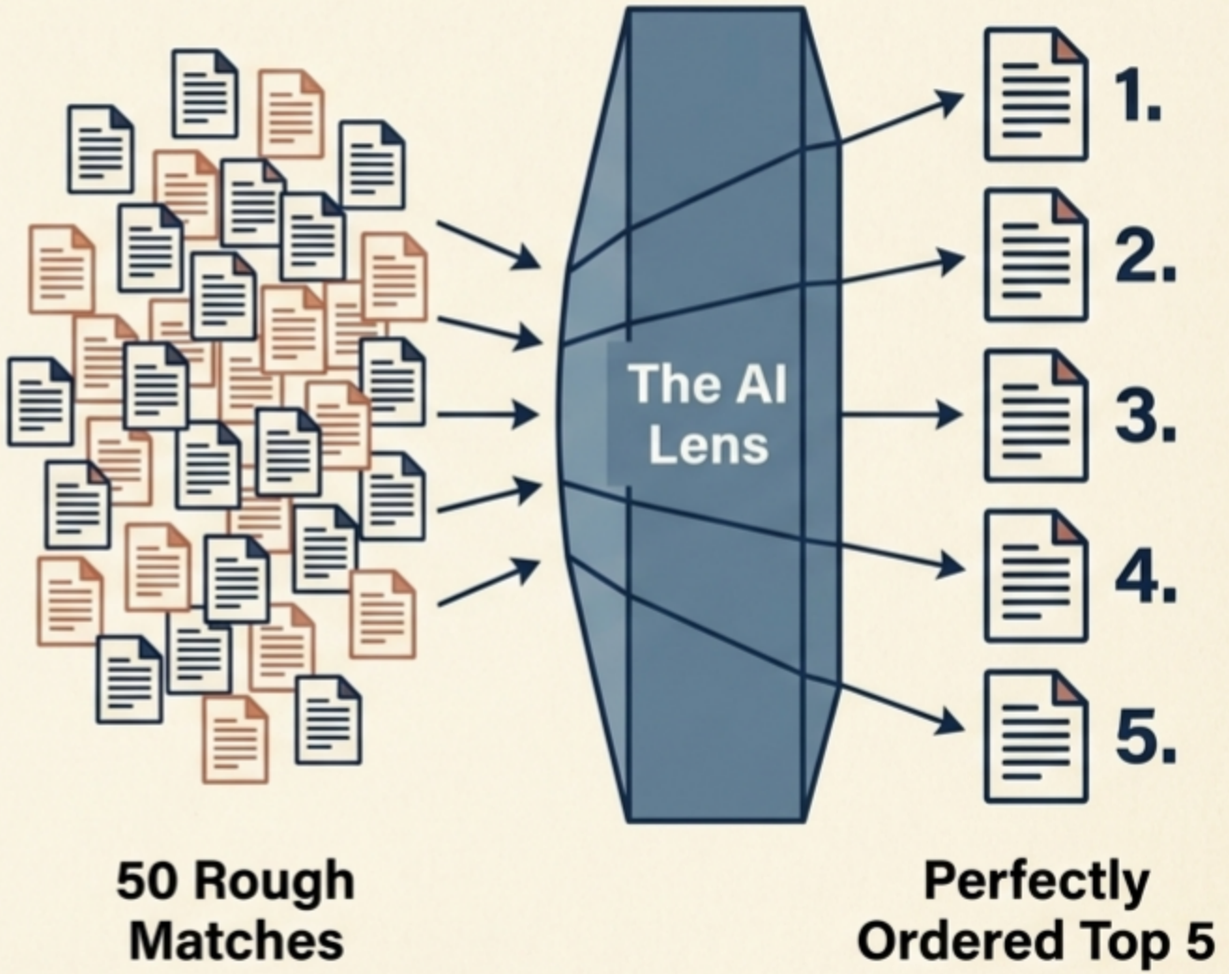


Step 3: Re-rank (Optional Precision)

Initial retrieval is fast but rough.

Re-ranking applies a slower, highly accurate model to **perfectly order the top candidates.**

Key Insight: Skip re-ranking only if first-stage results are sufficient. When precision matters, this step ensures the absolute best answer is placed at position #1.



Step 4: Generate (The RAG Synthesis)

Retrieval-Augmented Generation (RAG). An LLM reads the top passages retrieved by the search engine and synthesizes a grounded answer.

Key Insight: Skip generation if the user just needs a plain list of links. Use generation when the user needs an integrated answer built from multiple sources.

The Lexicon

Query

The human input or question driving the system.

Semantic Search

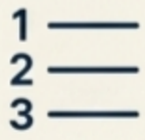
Searching by underlying meaning instead of literal spelling.

Embedding

Text translated into mathematical coordinates to map meaning.

Retrieval

Quickly extracting candidate documents from the whole corpus.



Ranking

Ordering retrieved documents from most to least relevant.

LLM

Large Language Model; the engine understanding and generating text.

RAG

Retrieval-Augmented Generation; using search results to write grounded answers.

Check Your Understanding

Q: In one sentence, why does keyword search miss relevant results?	**A:** It matches literal words, missing documents that express the exact same meaning with different vocabulary.
Q: What does Retrieve do, and how does ‘semantic’ change it?	**A:** It pulls candidate documents; ‘semantic’ means it compares meaning via embeddings instead of shared letters.
Q: Which pipeline stages are optional?	**A:** Re-rank (skip if initial results suffice) and Generate (skip for a traditional link-based search box).