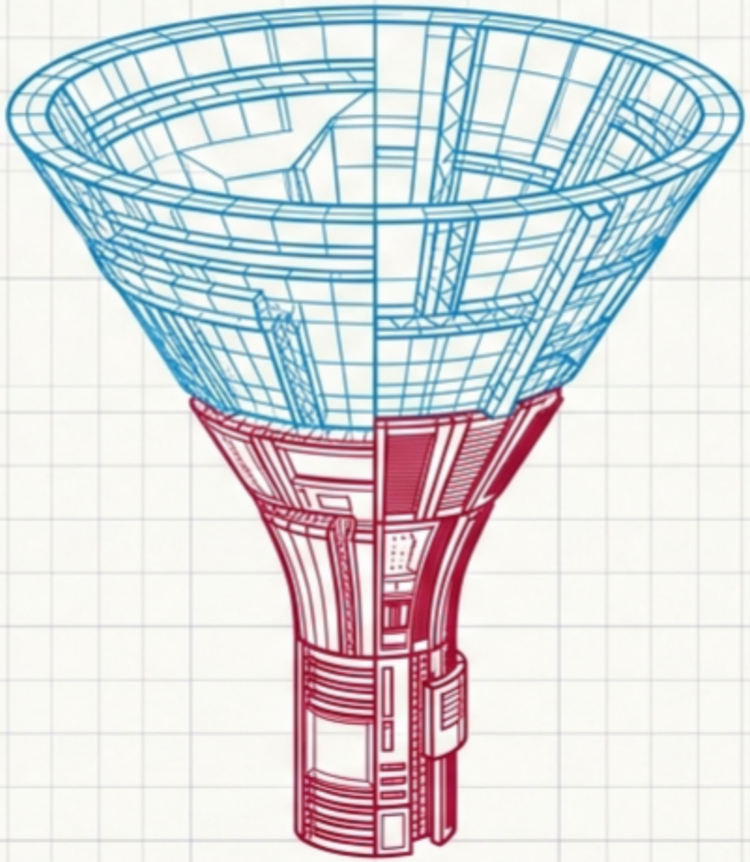
Semantic Search: The Retrieve-and-Re-Rank Pattern

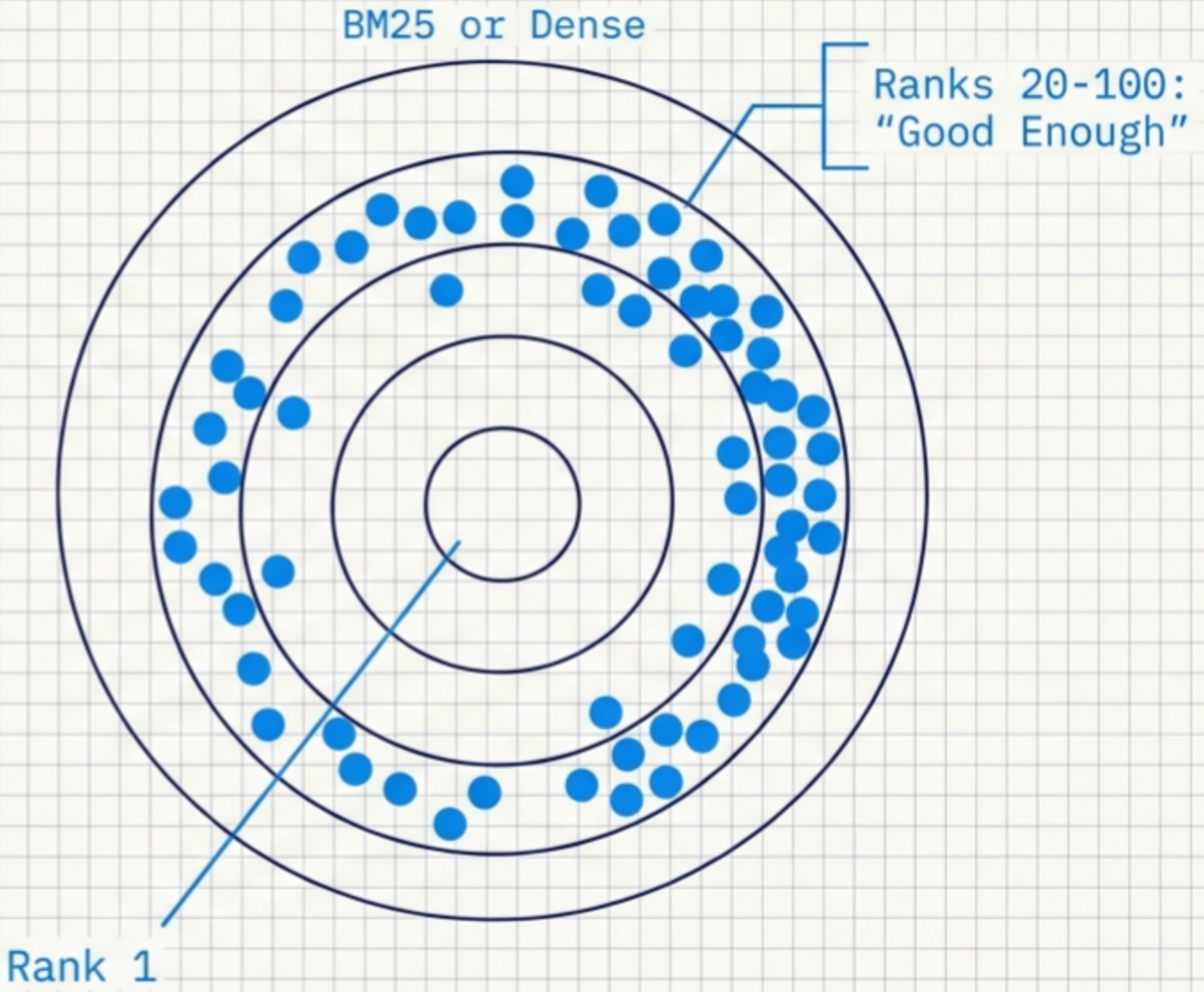
SYSTEM_ARCHITECTURE // MODULE 07



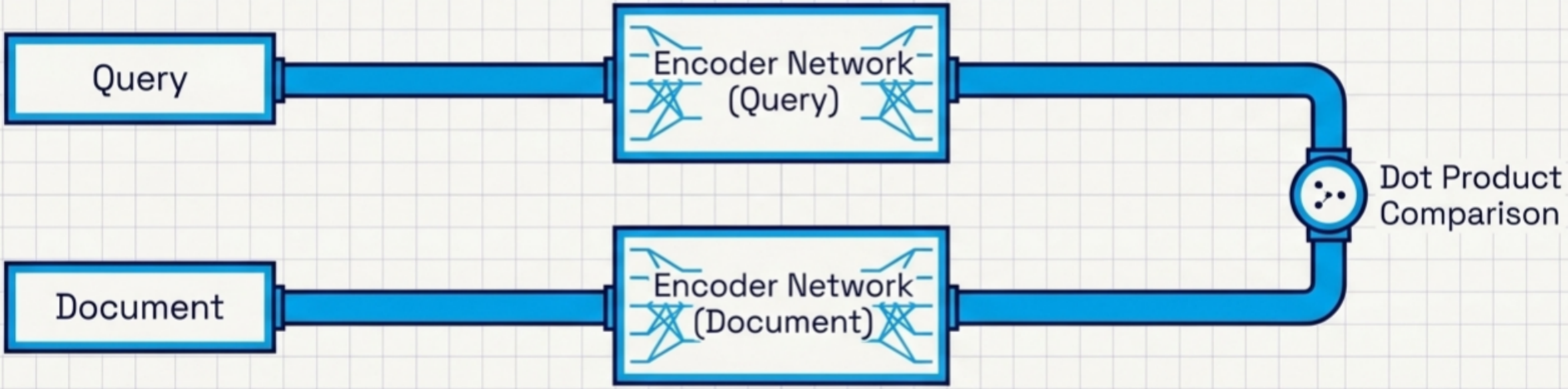
First-Stage Retrieval Gets You Close, Not Perfect

Fast retrieval algorithms prioritize speed over deep comprehension.

The genuinely “best” document usually lands in the top 20-100 results, but rarely surfaces exactly at Rank 1.

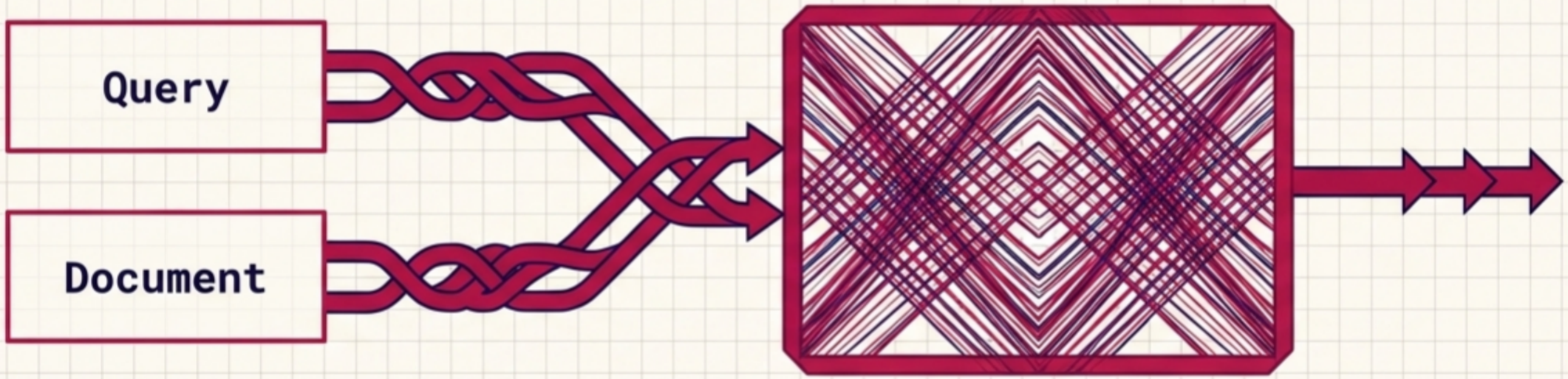


Bi-Encoders Embed Query and Document Separately



Advantage: Highly pre-computable. All documents can be embedded beforehand.	**Constraint:** Fast, but lacks deep contextual interaction between query and document.

Cross-Encoders Read Query and Document Togethen Together



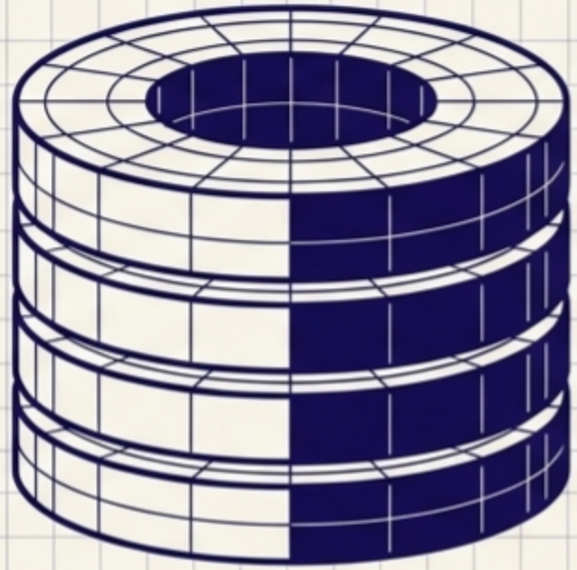
Advantage: Yields a far more accurate relevance score.

Constraint: Cannot pre-compute document embeddings. Must be scored live.

Matching the Encoder Architecture to the Task

	Bi-Encoder	Cross-Encoder
Architecture	Separate paths	Joint processing
Pre-computability	Yes (highly scalable)	No (requires live pairing)
Speed	Blazing fast	Compute heavy
Accuracy	Approximate matching	High-precision semantic scoring
Optimal Role	`First-Stage Retriever`	`Second-Stage Re-ranker`

The Compute Bottleneck of Cross-Encoding



Corpus size = N documents

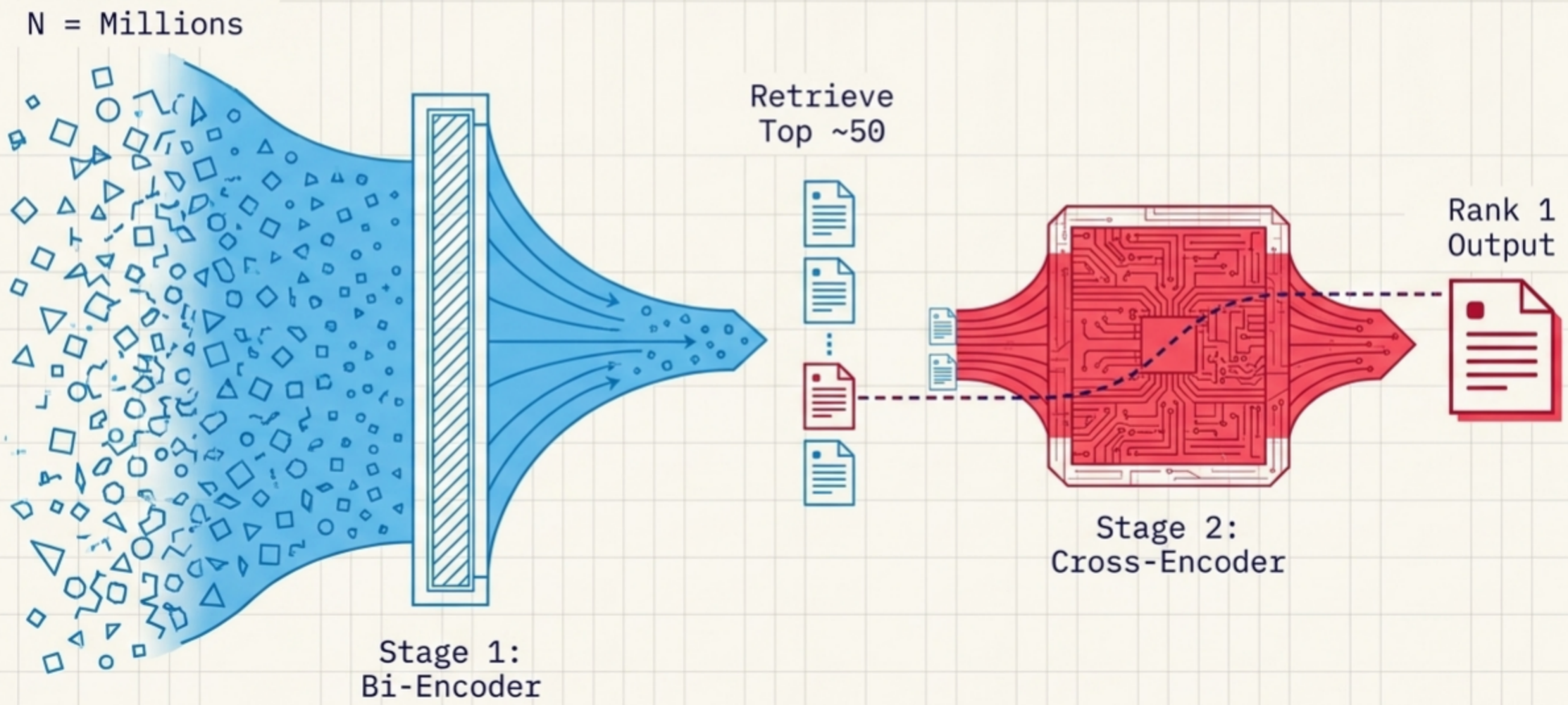
Requires

N model runs per query

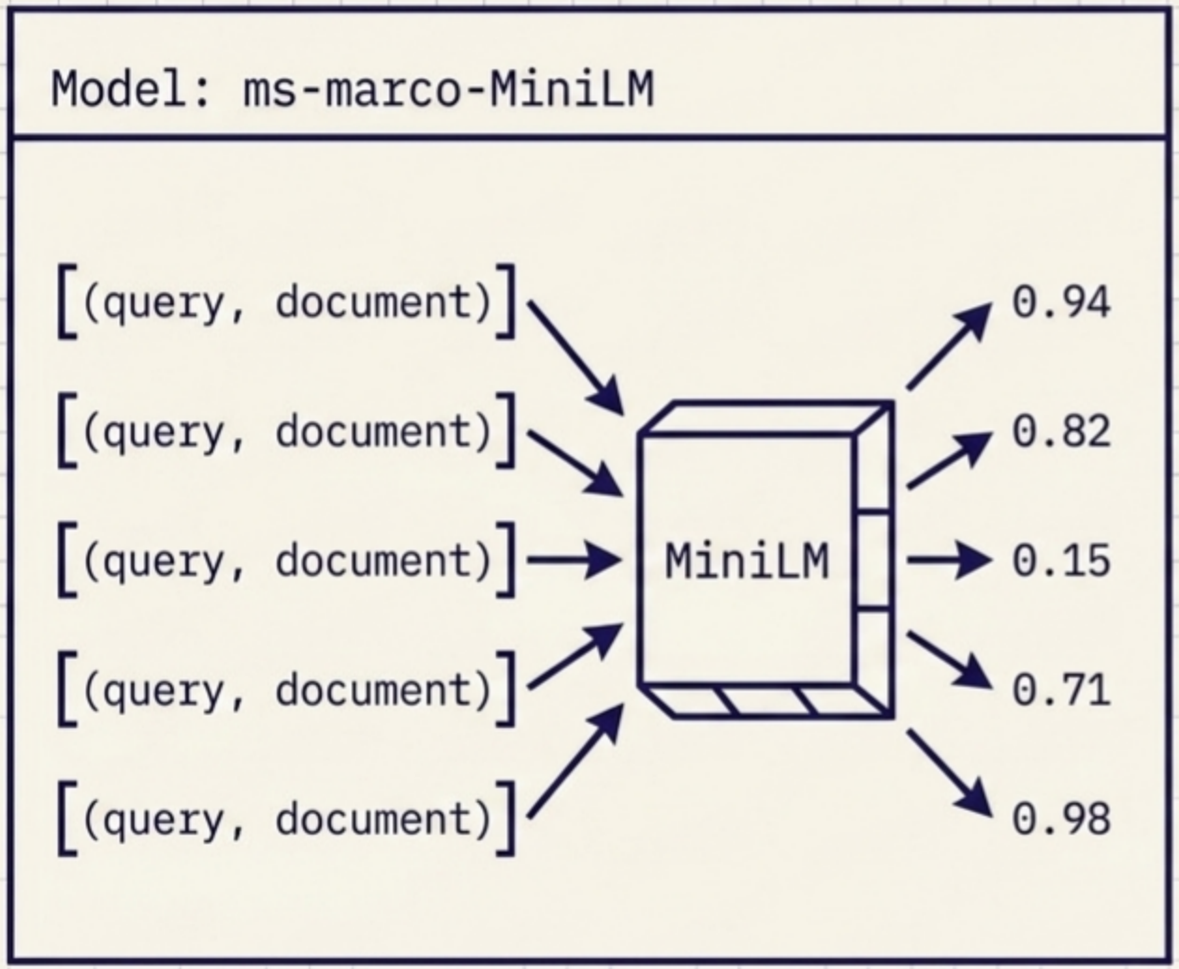
1%

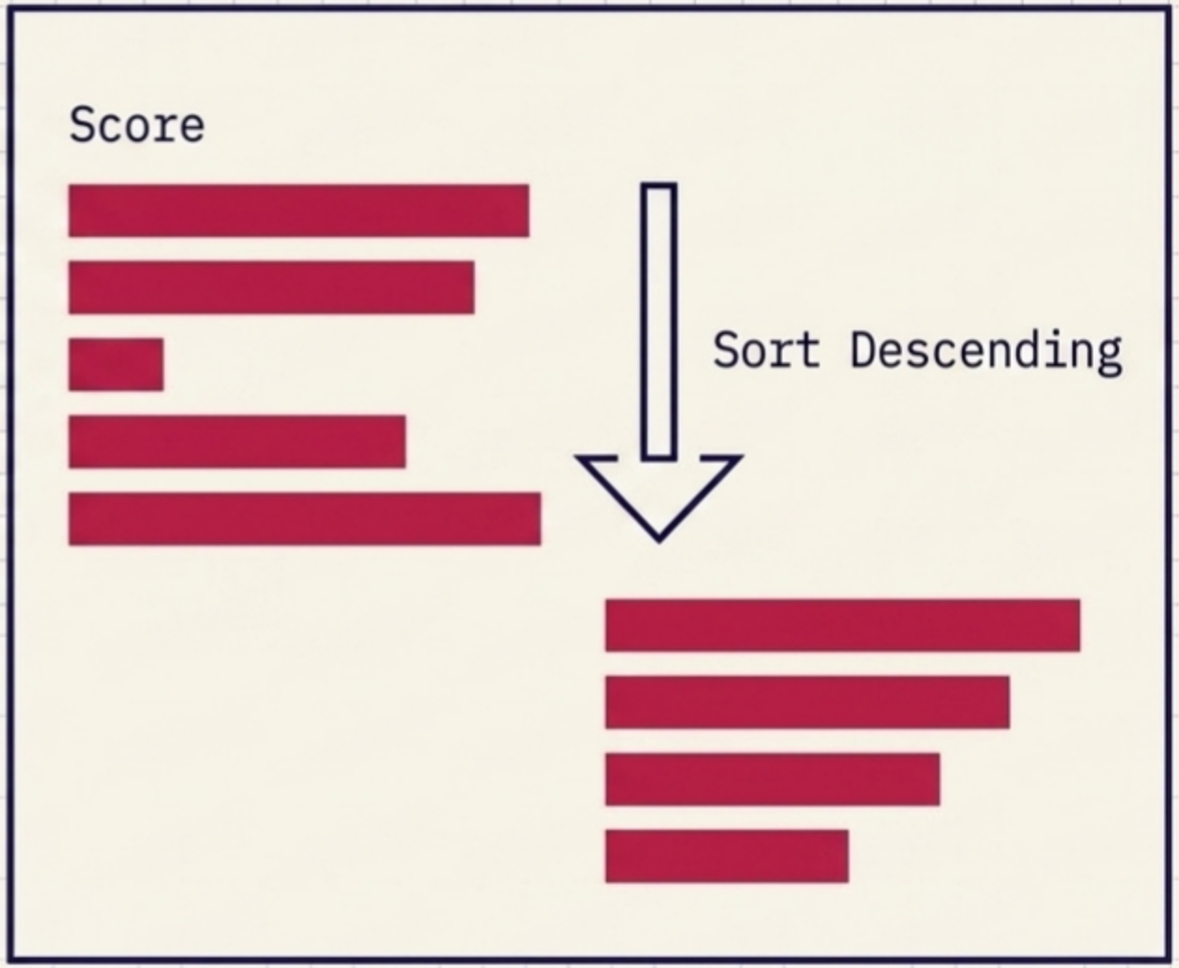
Running a cross-encoder against an entire corpus for every single query is hopeless at scale. We must severely limit the documents it processes.

Retrieve Wide and Cheap, Re-Rank Narrow and Precise

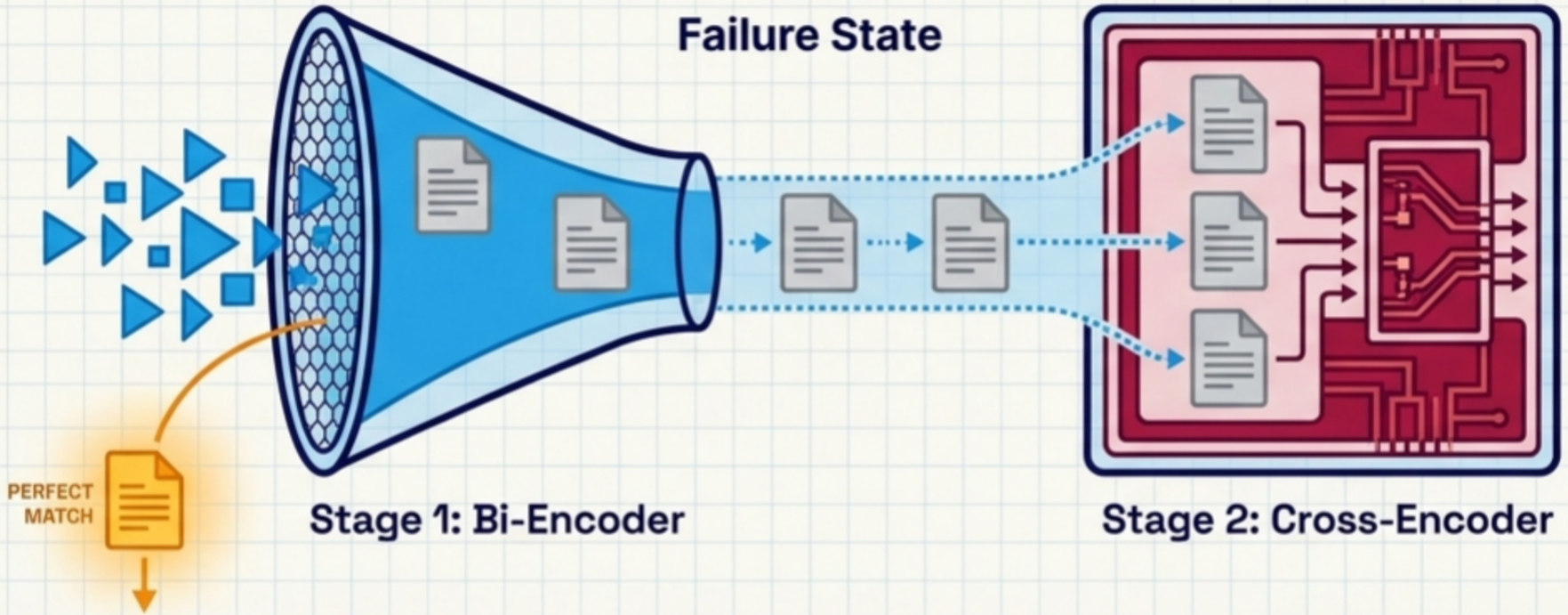


Deploying the Re-Ranker Model in Practice





A Re-Ranker Cannot Save Missed Documents



A re-ranker only reorders the candidate set it is given.

The Recall Bound: If the optimal document never makes the top-k in Stage 1, no re-ranker can retrieve it. You must fix first-stage recall first.

System Architecture Diagnostics

[re-ranking]

[cross-encoder]

[bi-encoder]

[two-stage retrieval]

[candidate set]

[first-stage recall]

What justifies the cost of a cross-encoder?

It reads query & document together for unmatched accuracy, though it cannot pre-compute.

Why not use cross-encoders at Stage 1?

Requires N model runs per query—impossible over a full corpus.

What is the ceiling of a re-ranker?

First-stage recall limits the system; you can only re-rank what the first stage catches.