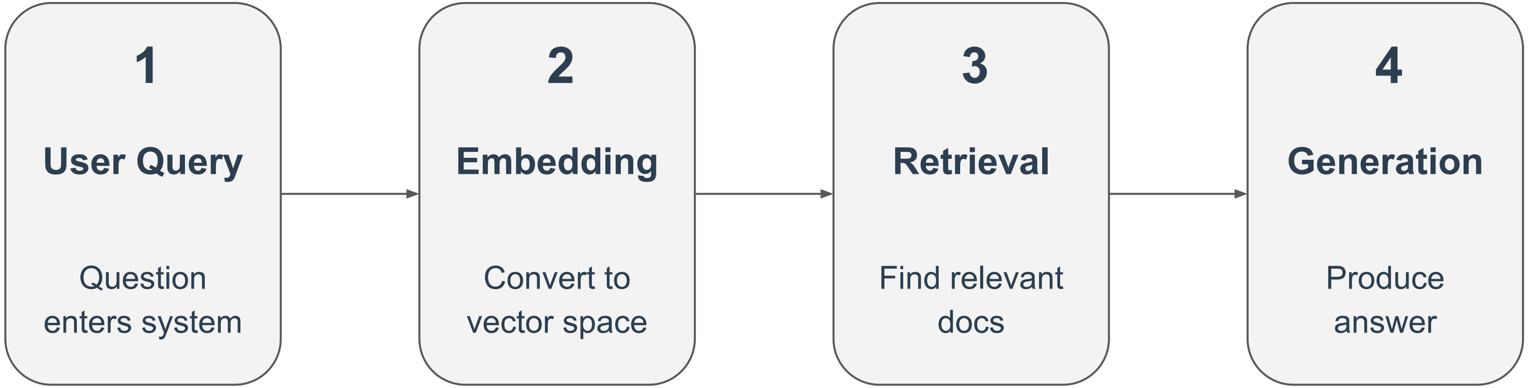
Robust Document Selection for RAG

An Optimization Approach

MIT MANAGEMENT
SLOAN SCHOOL

What is Retrieval-Augmented Generation?



What is Retrieval-Augmented Generation?

1
User Query
Question enters system

2
Embedding
Convert to vector space

3
Retrieval
Find relevant docs

4
Generation
Produce answer

Retrieval quality directly shapes answer quality

Top-K Retrieval Looks Simple — But Breaks Easily

Standard Top-K retrieval selects documents with the highest similarity scores

redundant context, low diversity, and unstable results when embeddings are noisy

Top-K Retrieval Looks Simple — But Breaks Easily

Standard Top-K retrieval selects documents with the highest similarity scores

redundant context, low diversity, and unstable results when embeddings are noisy

Better retrieval requires more than ranking

Our Approach:

Formulated as a robust optimization problem

1. Balance relevance and conciseness

pick docs that add the most value; penalize extra docs to avoid noise

$$
\begin{aligned}
\max_{x,y} \; \min_{\{\tilde{\mu}_i \in \mathcal{U}_i\}_{i\in I}} \quad & \sum_{i\in I} s_i(\tilde{\mu}_i)\, x_i \;-\; \lambda \sum_{i\in I} x_i \\
\text{s.t.} \quad & y_{ij} \le x_i x_j, & \forall i<j \\
& \max_{\tilde{\mu}_i \in \mathcal{U}_i,\, \tilde{\mu}_j \in \mathcal{U}_j} \sum_{i<j} \cos_{ij}(\tilde{\mu}_i, \tilde{\mu}_j)\, y_{ij} \le \rho_{\mathrm{div}} \sum_{i<j} y_{ij}, \\
& x_i \in \{0,1\}, \quad y_{ij} \in [0,1], & \forall i \in I,\, \forall i<j
\end{aligned}
$$

2. Stress-test embeddings

Account for uncertainty and adversarial perturbations

3. Encourage diversity

Experimental Setup

Mini-Wikipedia RAG dataset

3,200 passages, 918 factual QA pairs

Factual verification and entity-based queries

Embedding models tested

BERT

Instructor

E5

MPNet

What We Found

Robust optimization matches or outperforms Top-K retrieval

retrieval accuracy and consistency ⬆

True similarity depends on a small set of stable embeddingdimensions

This creates a cleaner foundation for downstream generation