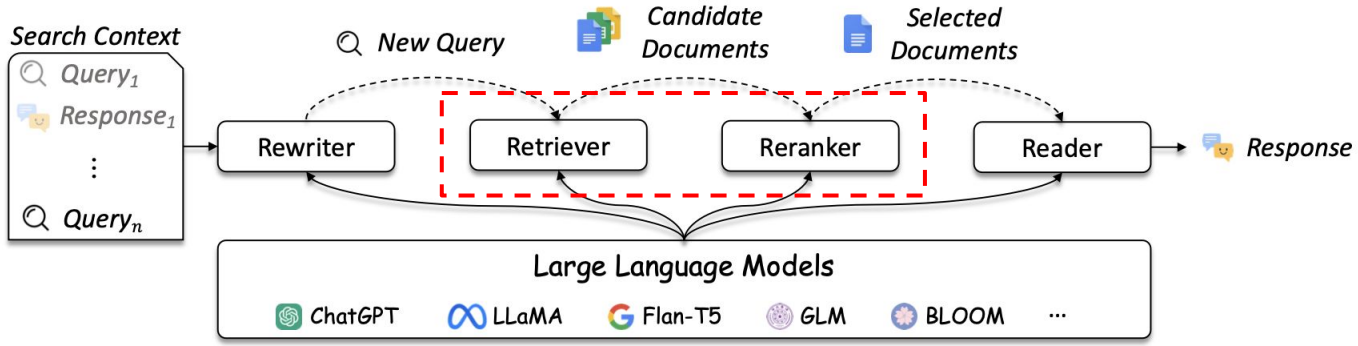
RAG systems are powerful yet imperfect



Retriever

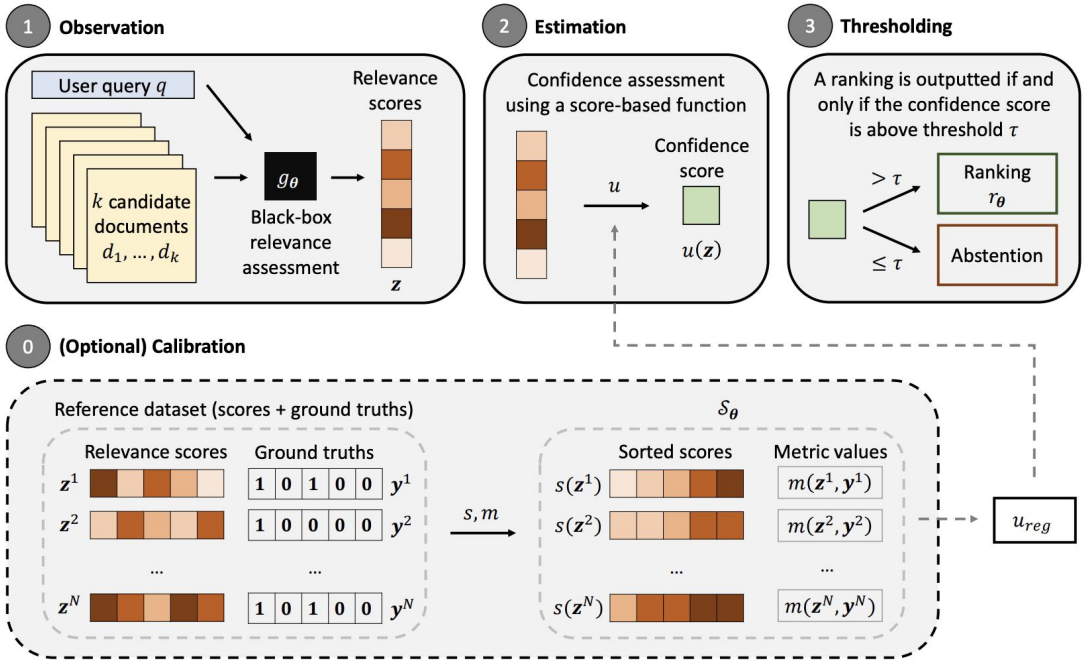
> **Input**: query + document database
> **Output**: retrieved documents
> **Objective**: efficient selection of relevant documents, with high recall
> **Method**: frequentist or LLM-based (bi-encoder)

Reranker

> **Input**: query + retrieved documents
> **Output**: final document selection
> **Objective**: finer document selection, with high precision (focus on effectiveness)
> **Method**: LLM-based (bi- or cross-encoder)

Risks: poor encoding, incomplete database, unclear query, poor retrieval…
⇒ IRRELEVANT ANSWER!

It’s better to abstain when you’re not sure!



> **Context**: reranking stage
> **Constraints**: black-box model access, low computational overhead
> **Methods**: heuristic- (blue) and reference-based (green)
> **Evaluation**: normalized area under the metric-abstention curve averaged over 6 datasets and 22 models

	max	std	p1-p2	linreg
mAP	27.5	28.5	21.5	**37.4**
mNDCG	29.4	29.1	22.4	**38.0**
mRR	34.6	32.3	28.9	**40.7**

Goal 1: show that abstention works (increasing metric vs. abstention curve, i.e., positive normalized AUC)

Goal 2: show that access to a reference set improves abstention effectiveness