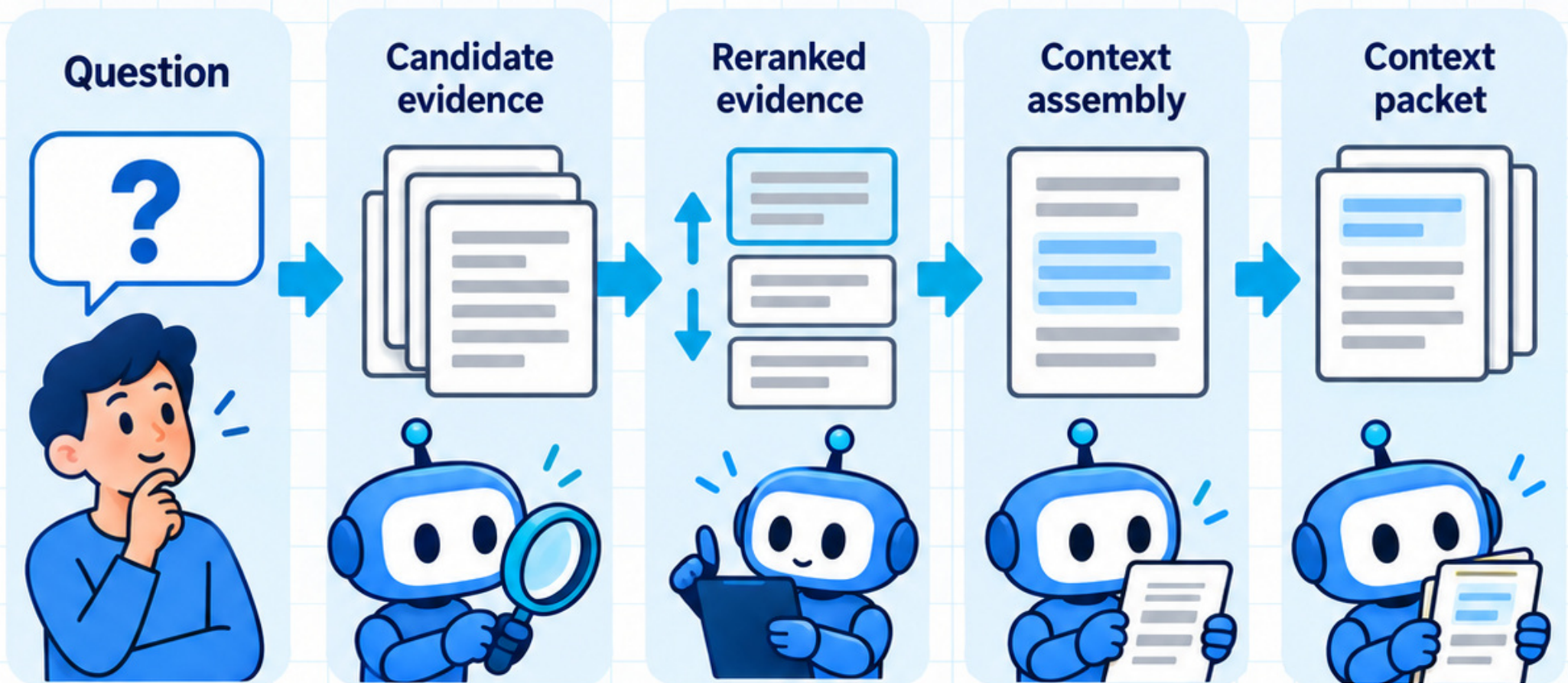


RETRIEVE AND RERANK EVIDENCE

Retrieval selects candidate evidence. Reranking reorders those candidates for the question.

Context assembly decides what the model actually receives.

Today: design each step for a fictional return-policy assistant.



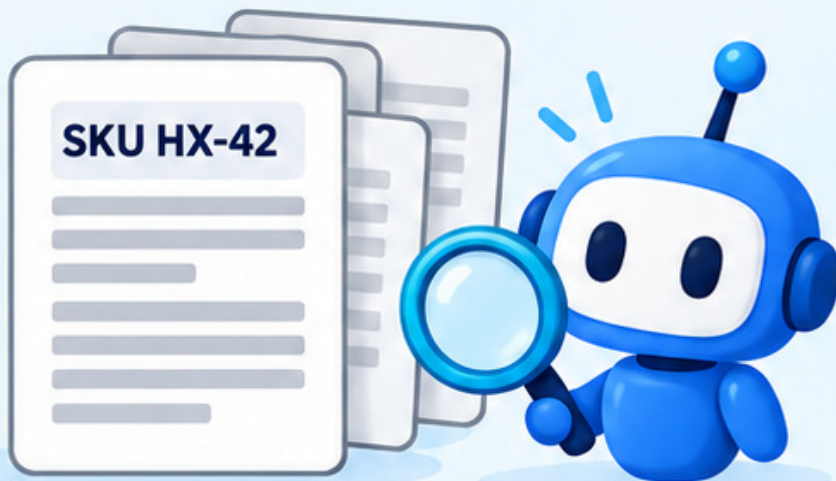
TWO QUESTIONS NEED DIFFERENT CLUES



Toy question A:
What is the return rule
for SKU HX-42?



Identifier search



Toy question B:
These headphones hurt.
Can I send them back?



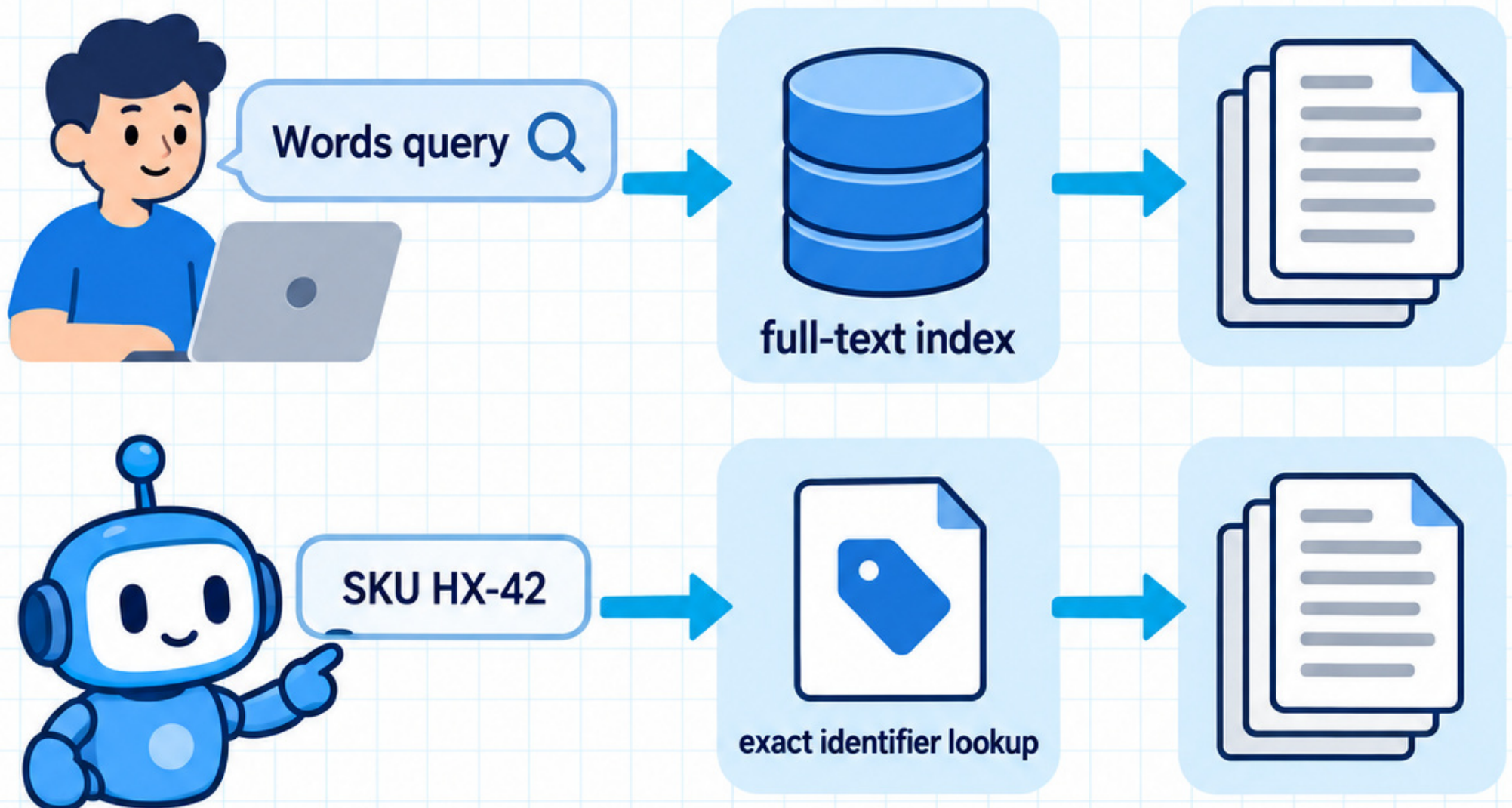
Paraphrase search



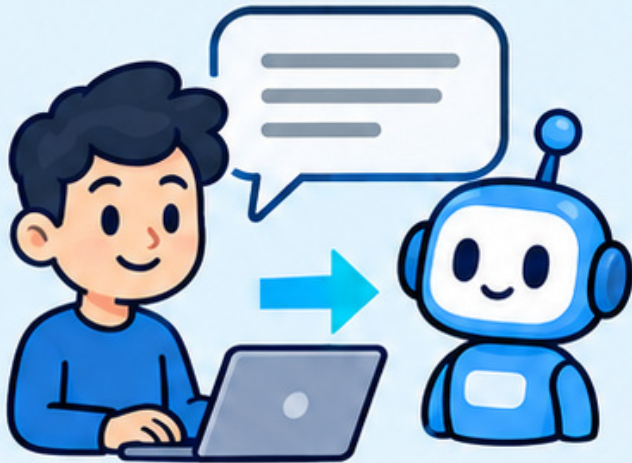
**Exact identifiers and paraphrased meaning
can need different retrieval signals.**

LEXICAL SEARCH FINDS TERMS

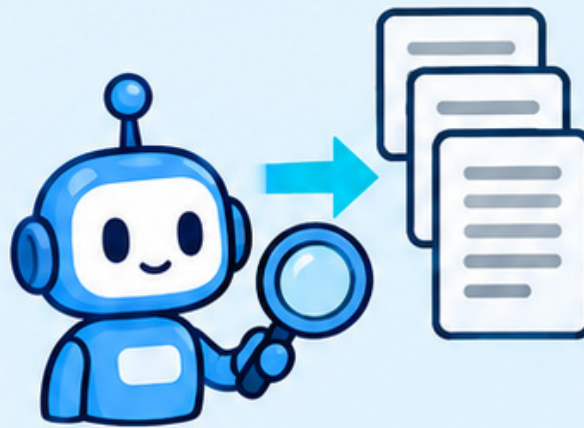
Lexical search matches indexed terms in the query and documents. PostgreSQL full-text search supports token processing and ranking. For strict product IDs, use an explicit identifier field or exact-match path.



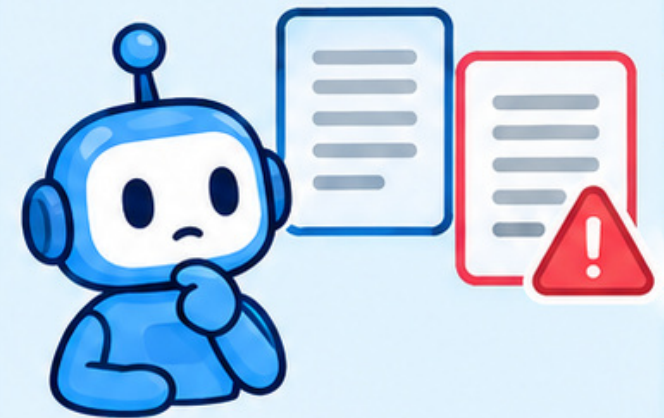
VECTOR SEARCH FINDS NEIGHBORS



Encode the query with the compatible embedding contract.



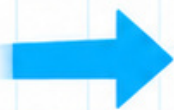
Find nearby document vectors under the chosen metric.



Semantic similarity can help with paraphrases, but can confuse similar-looking policies.



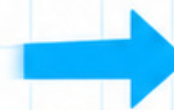
Paraphrase



Query embedding



Nearby policy vectors



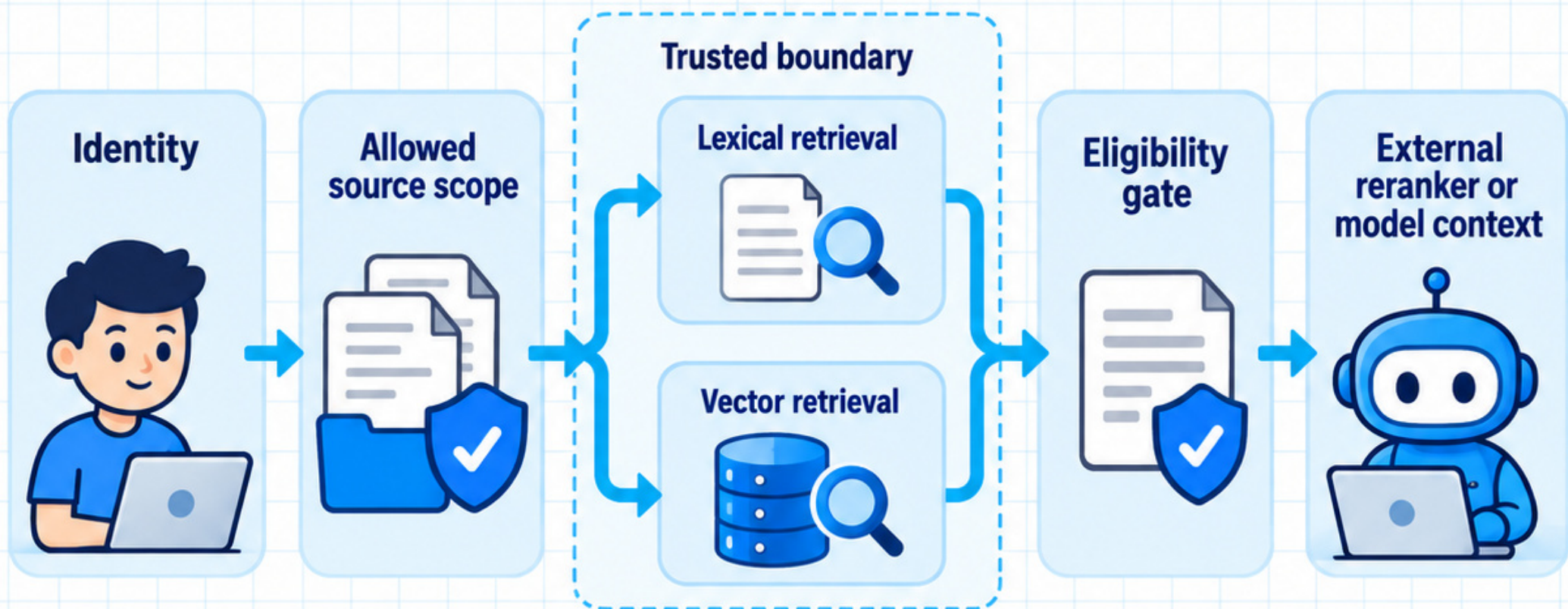
Relevance check

START WITH ELIGIBLE EVIDENCE

Resolve identity and the allowed source scope first.

Enforce permissions, active versions and validity for both retrieval branches.

No unauthorized text may enter an external reranker or model context.

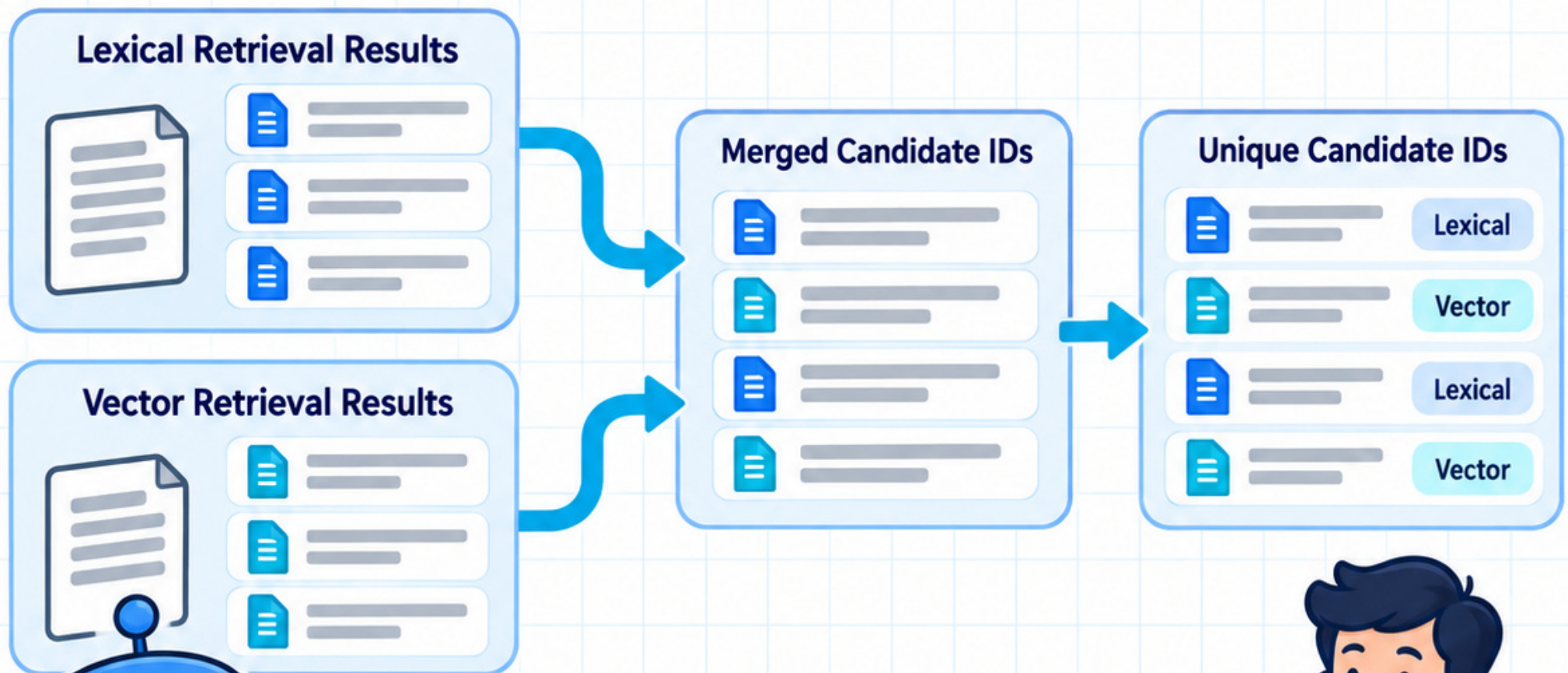


HYBRID COMBINES CANDIDATES

Run lexical and vector retrieval over the allowed scope.

Merge their candidate IDs and remove exact duplicates.

Keep source and score provenance so you can explain which branch supplied each result.



SCORES ARE NOT INTERCHANGEABLE

A lexical score and a vector distance have different meanings.
Do not add them without a justified calibration rule.
Rank fusion is one way to combine orderings without treating raw scores as equal units.

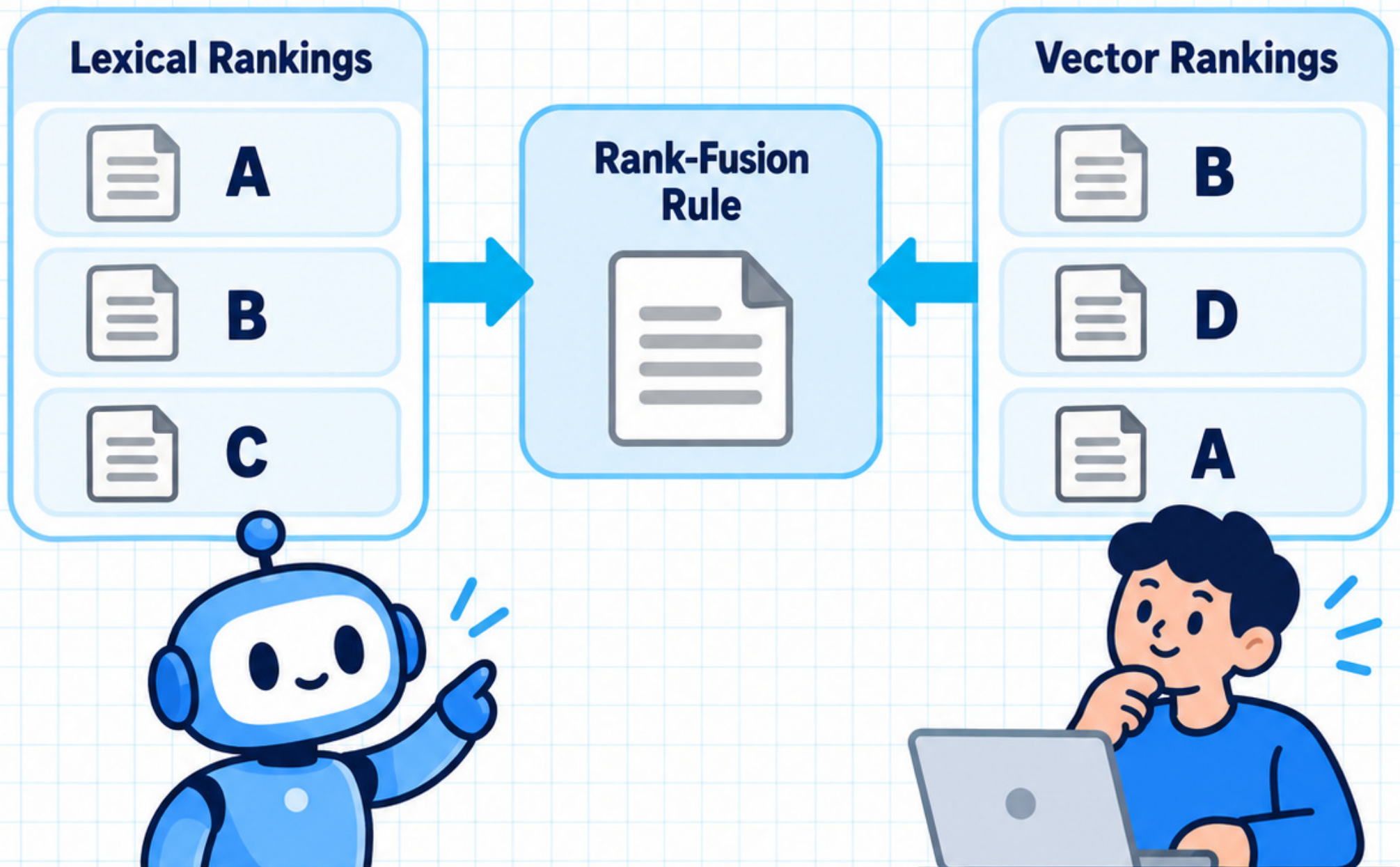


A SMALL RANK-FUSION EXAMPLE

Toy rankings: lexical = A, B, C; vector = B, D, A.

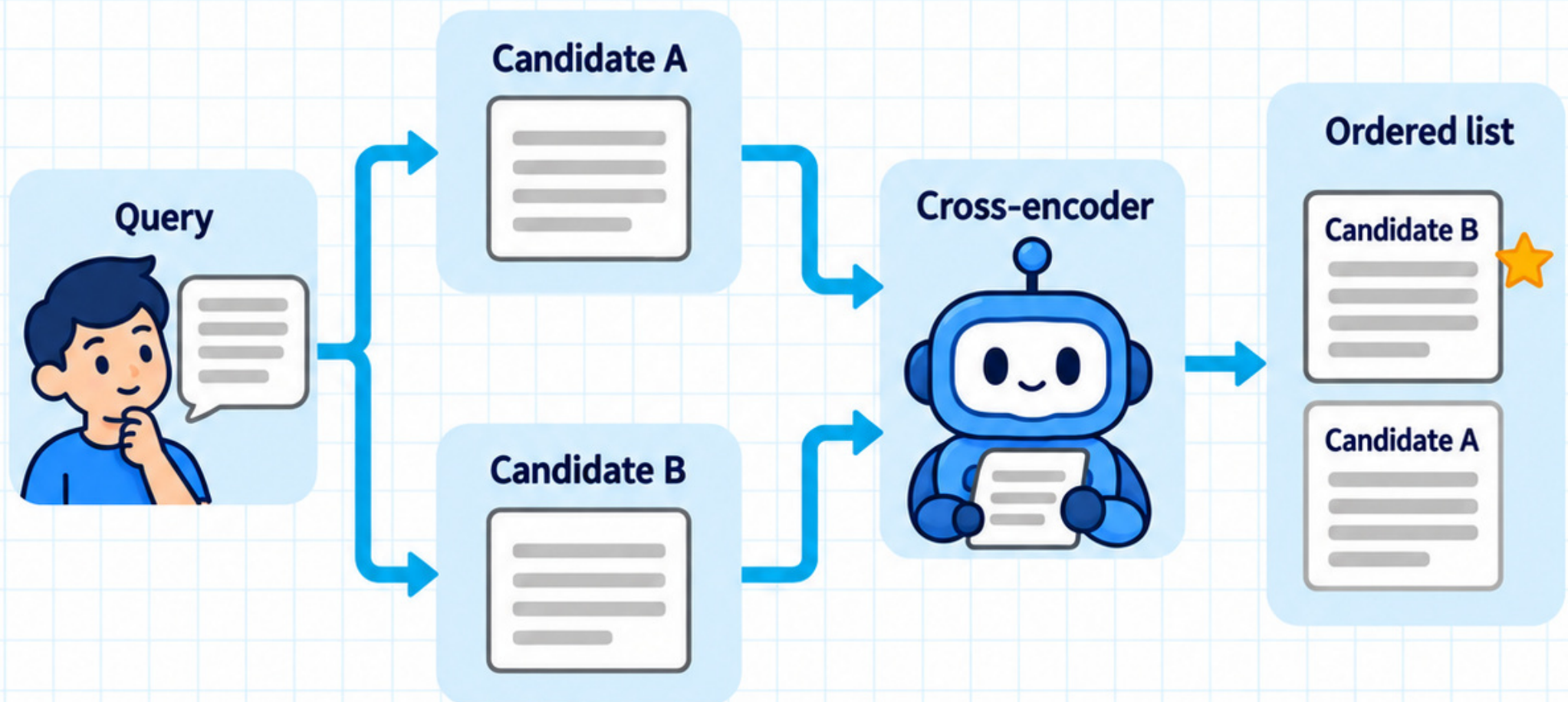
A and B appear in both lists; C and D appear in one.

A rank-fusion rule combines positions. Declare its formula and tie handling.



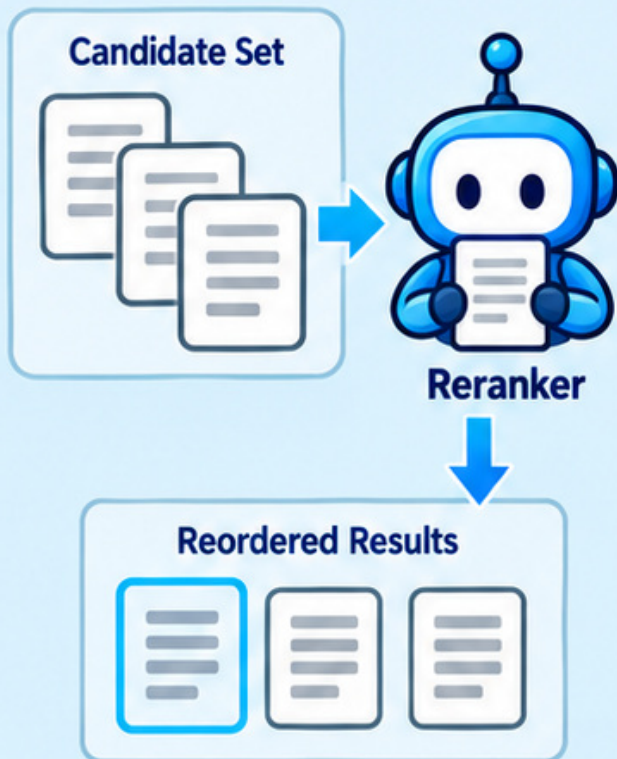
RERANK QUERY AND PASSAGE

A cross-encoder can score the query together with each candidate passage. It uses a richer pairwise comparison than a stored vector lookup. It adds computation, so evaluate whether the changed order helps.

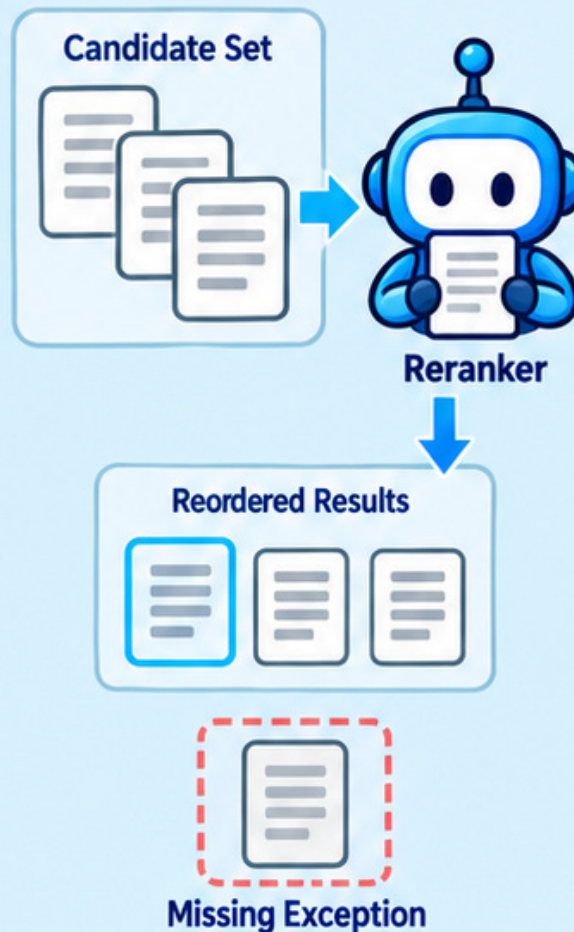


RERANKING CANNOT FIND MISSING TEXT

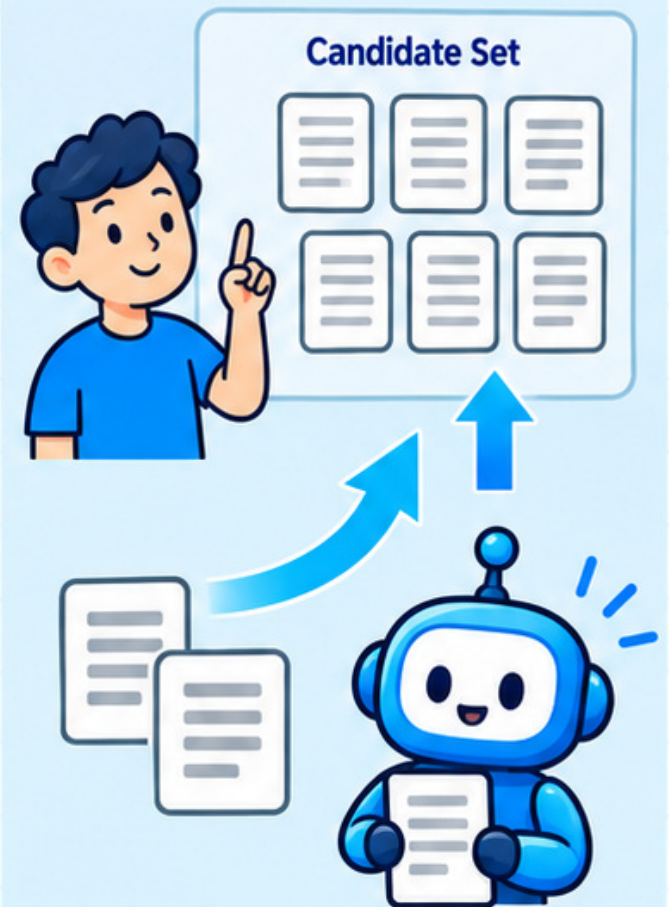
A reranker can reorder only the candidates it receives.



If the exception never entered the candidate set, reranking cannot recover it.



Improve candidate coverage before tuning the final order.

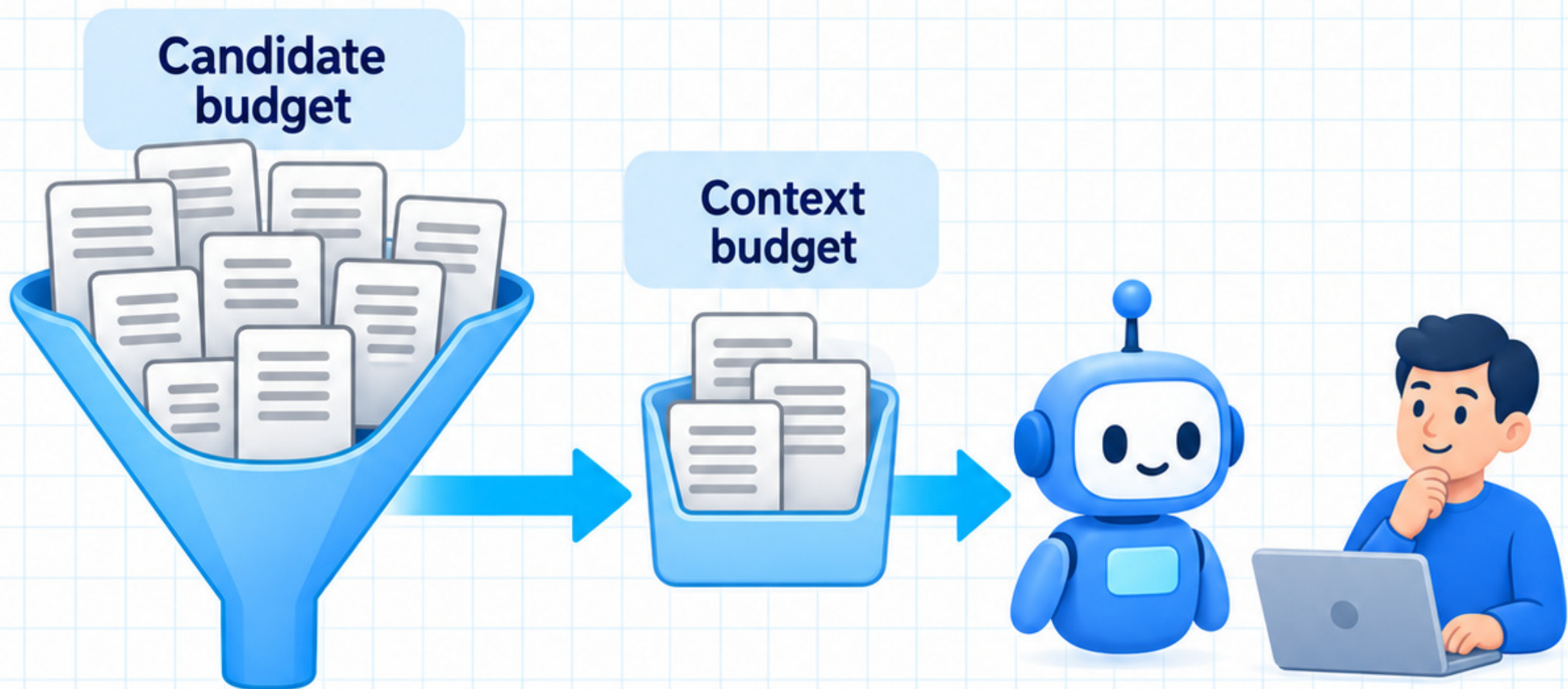


SEPARATE THE TWO BUDGETS

The candidate budget limits how many passages are considered.

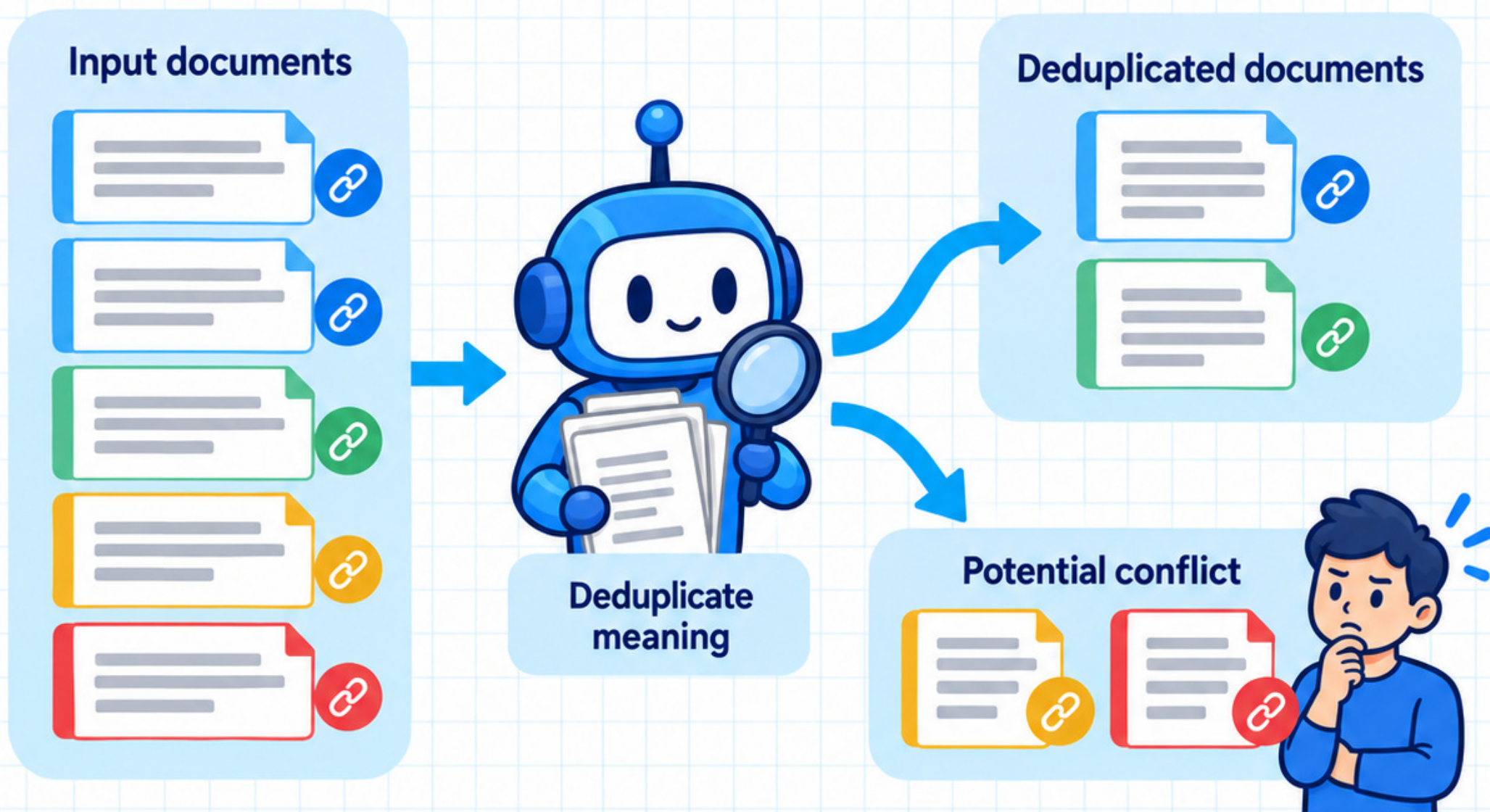
The context budget limits the text sent to the model.

A larger candidate pool can improve coverage while also adding retrieval or reranking cost.



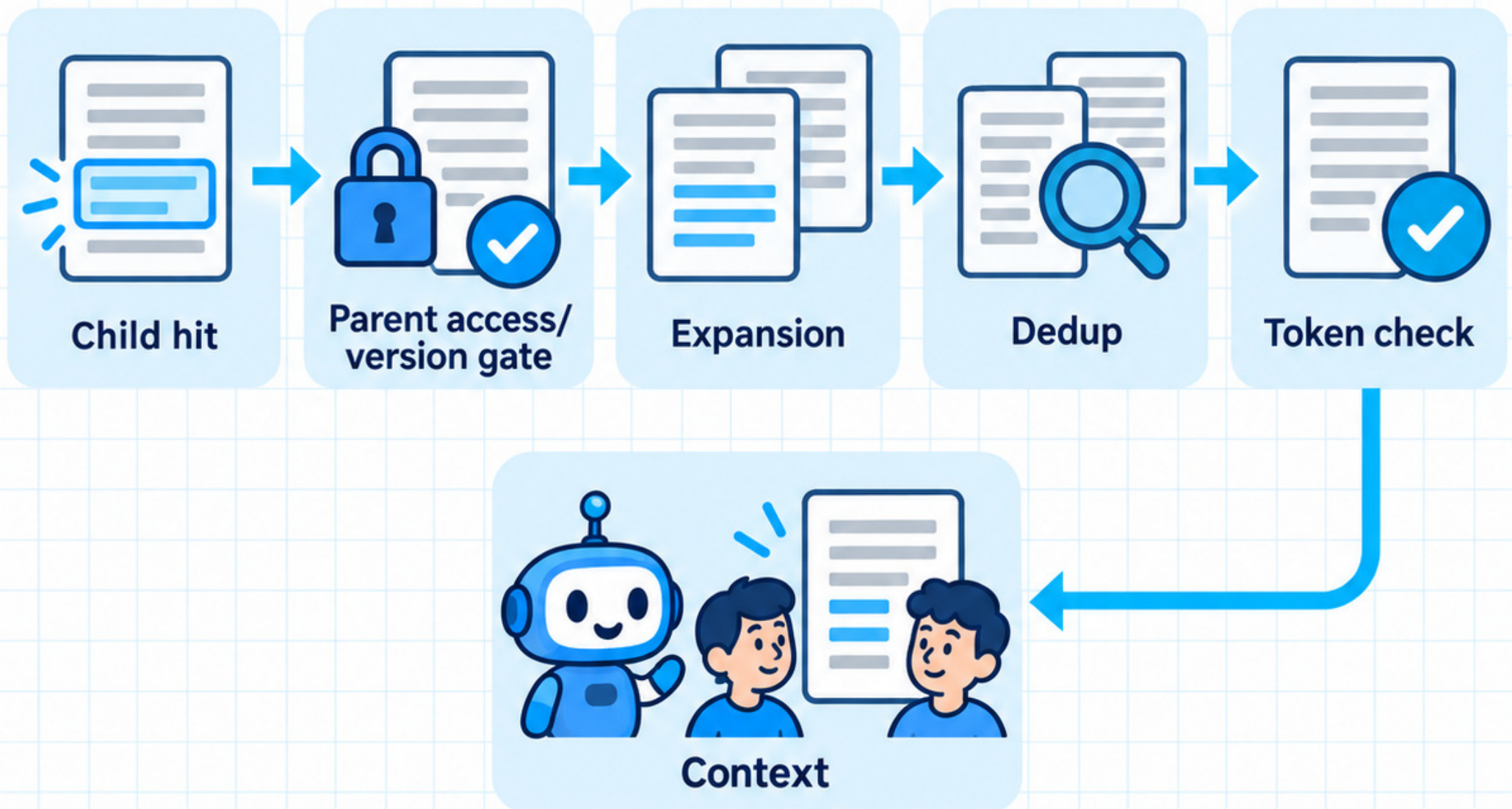
DEDUPLICATE MEANING CAREFULLY

Remove repeated copies that waste context space.
Retain source links and necessary rule-exception pairs.
Similar text from different valid authorities may reveal a conflict, so do not silently discard it.



EXPAND WITH A PURPOSE

- Fetch an authorized parent section when a child needs context.
- Keep the applicable source version and recheck access.
- After expansion, rerun deduplication and token-budget checks.



ASSEMBLE THE EVIDENCE PACKET



Give each selected passage a stable reference ID.



Include its source, version, location and relevant text.



Keep evidence separate from instructions; retrieved documents cannot grant new permissions.

Retrieved Documents



Evidence Packet



Reference ID



Source



Version



Location



Relevant Text

Instruction Panel



Instruction Panel content area with horizontal lines representing text.



HANDLE WEAK OR EMPTY EVIDENCE

If evidence is missing, ask for needed context or return a defined no-evidence response.

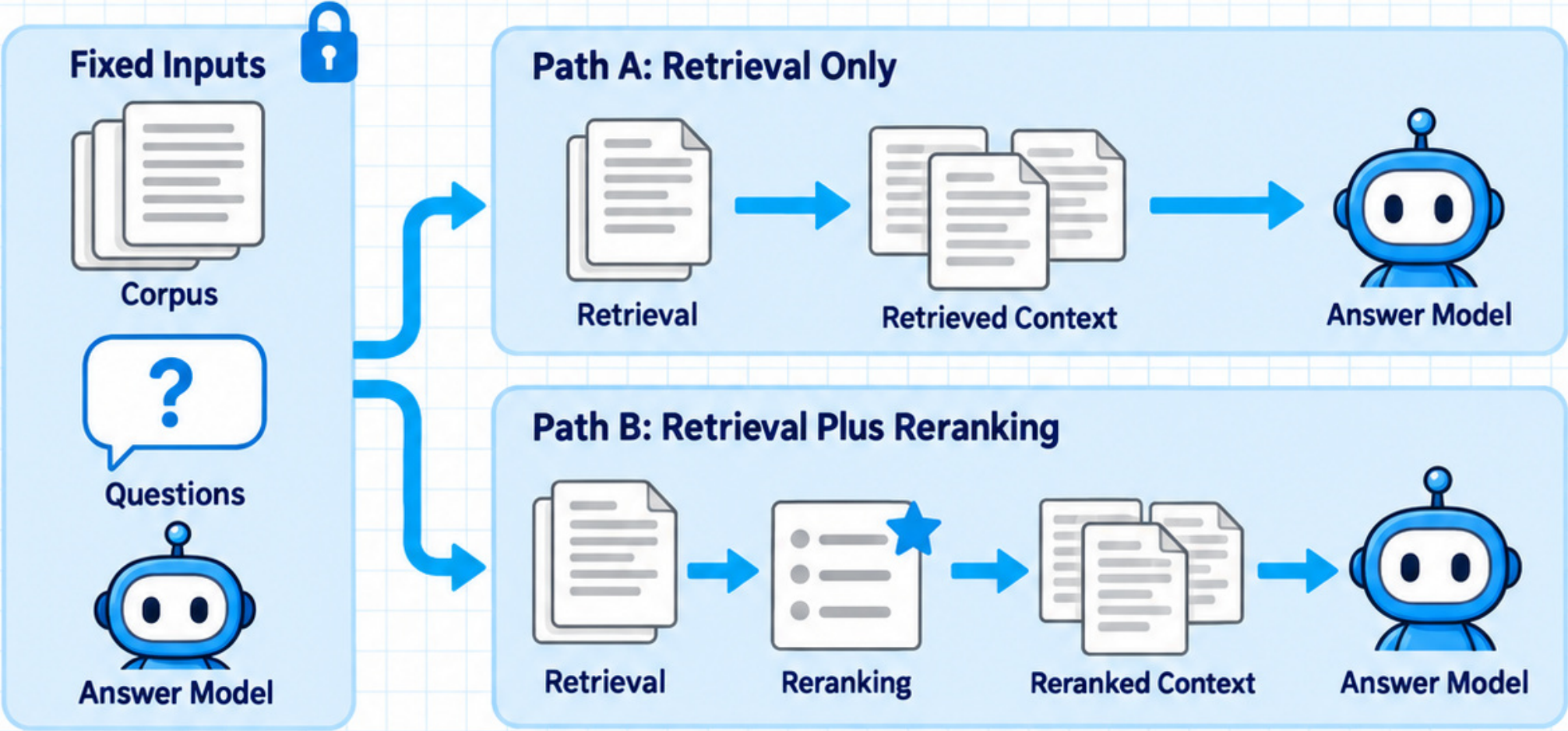
If sources conflict, show the conflict or escalate under policy.

Do not turn a low-quality retrieval result into a confident claim.



COMPARE TWO COMPLETE PATHS

Keep the corpus, questions and answer model fixed.
Compare retrieval alone with retrieval plus reranking.
Report evidence relevance, answer quality, context tokens and end-to-end latency.



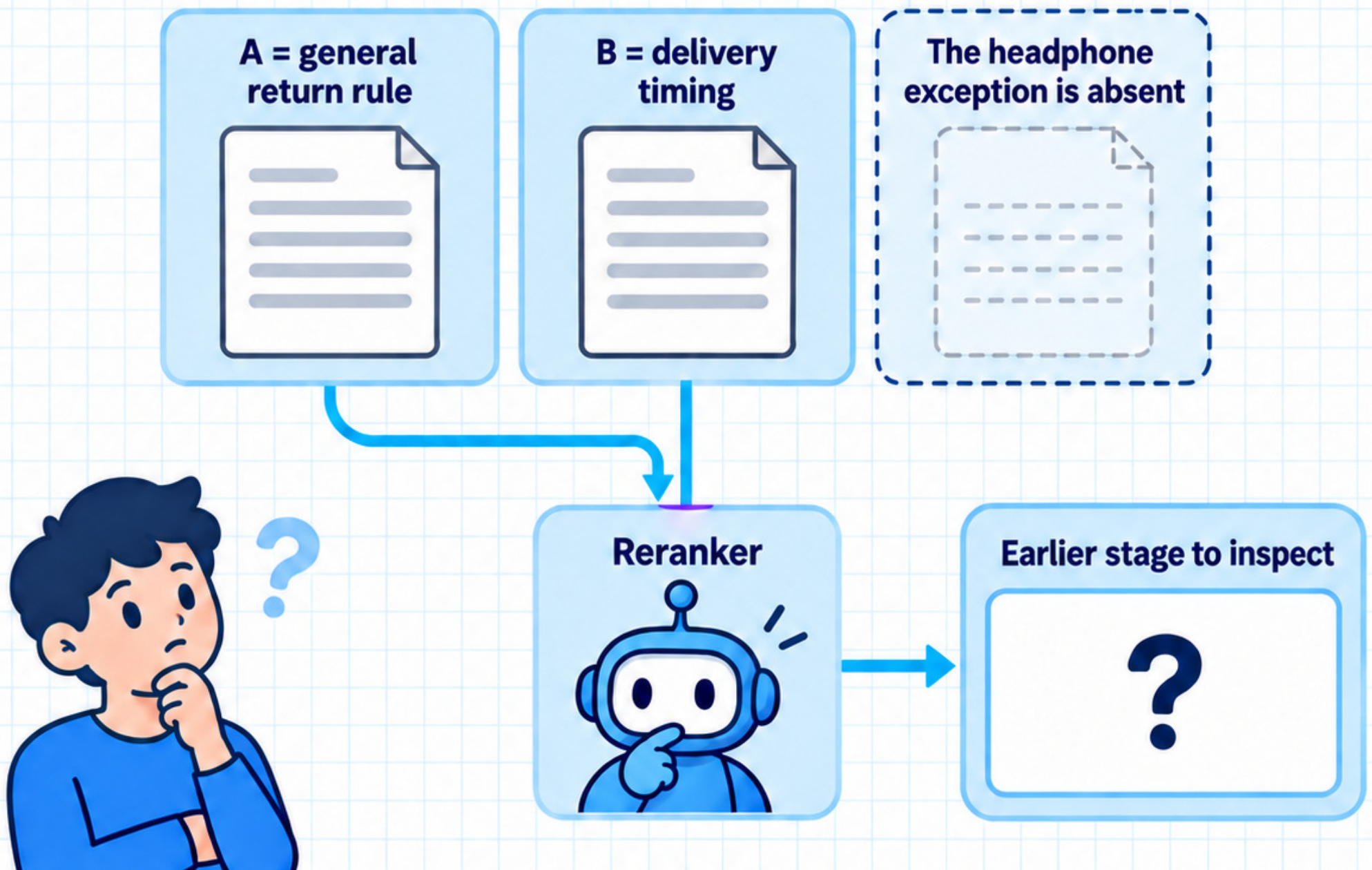
Path	Evidence Relevance	Answer Quality	Context Tokens	End-to-End Latency
Retrieval Only				
Retrieval Plus Reranking				

TRY THE MISSING-EXCEPTION CASE

Toy candidates: A = general return rule; B = delivery timing.

The headphone exception is absent.

Will adding a reranker fix the answer? Name the earlier stage you would inspect.



YOUR CANDIDATE-TO-CONTEXT DESIGN

Choose signals for the actual questions.

Measure candidate coverage before reranking. Enforce access before any external text use.

Next: connect authorized evidence to supported claims and citations.

