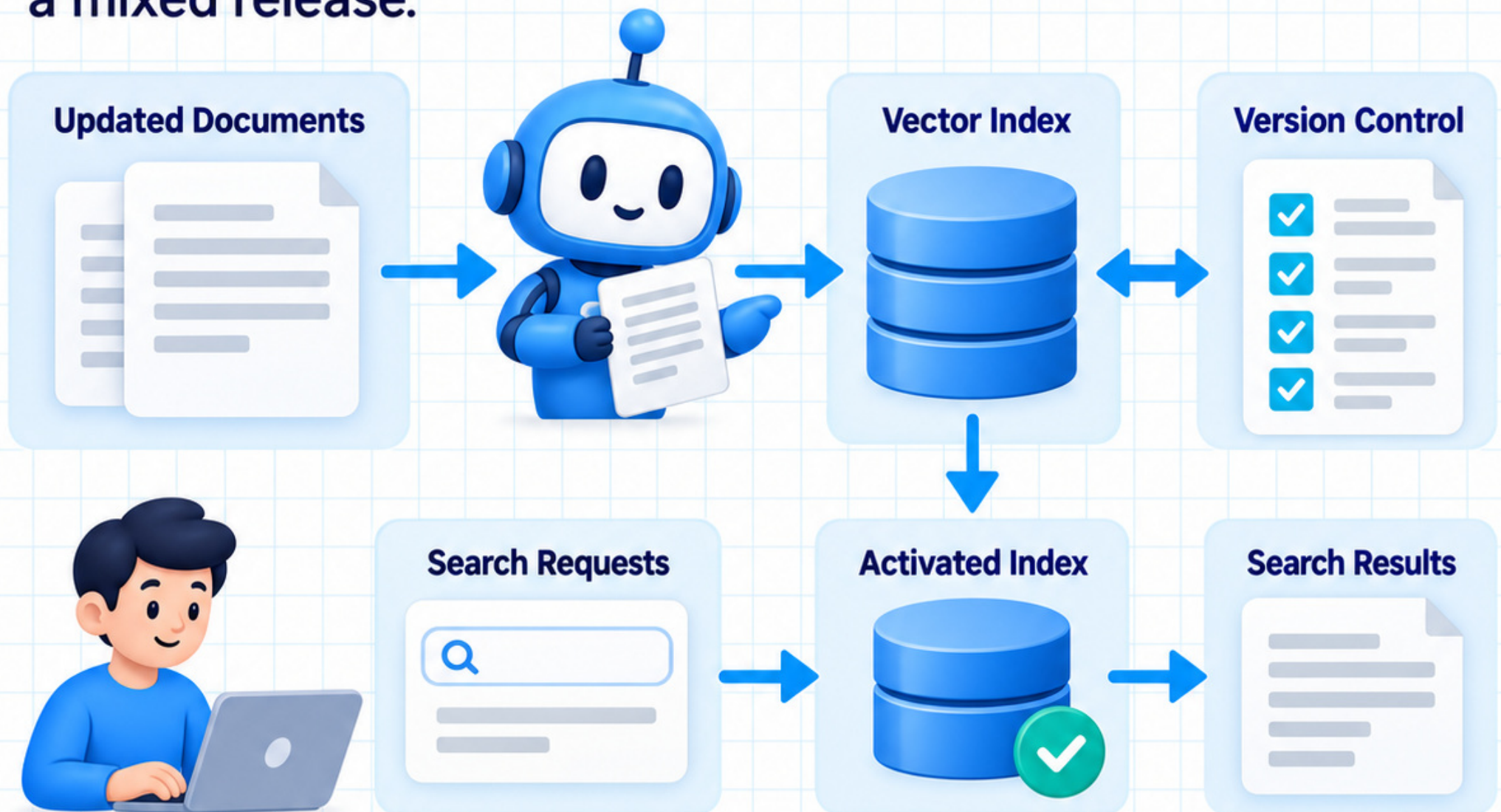


BUILD AND UPDATE THE INDEX

A vector index helps find nearby embeddings.

A production index also needs a contract for versions, permissions, activation and deletion.

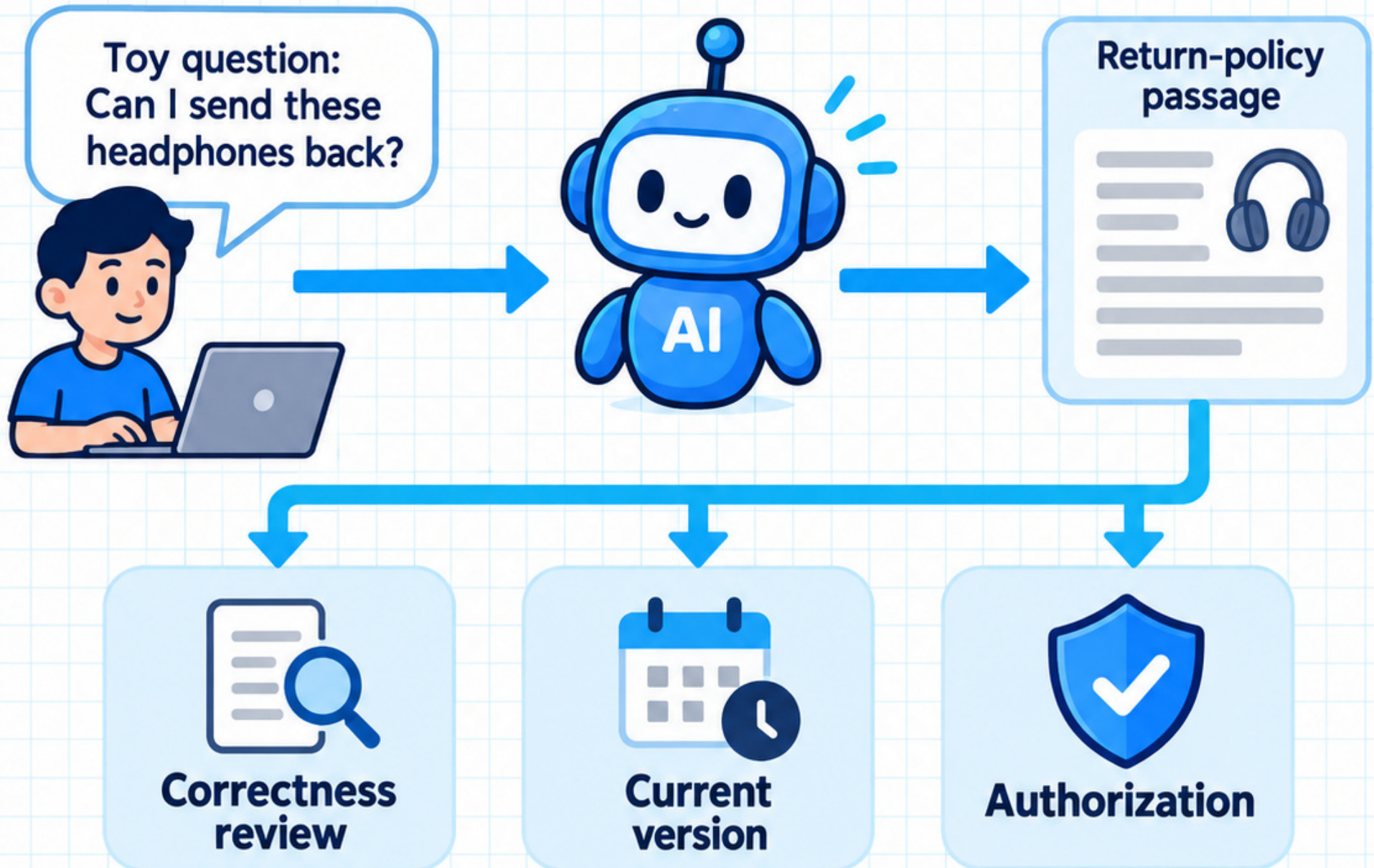
Today: update a fictional return policy without serving a mixed release.



SIMILARITY IS ONE SIGNAL

Toy question: Can I send these headphones back?

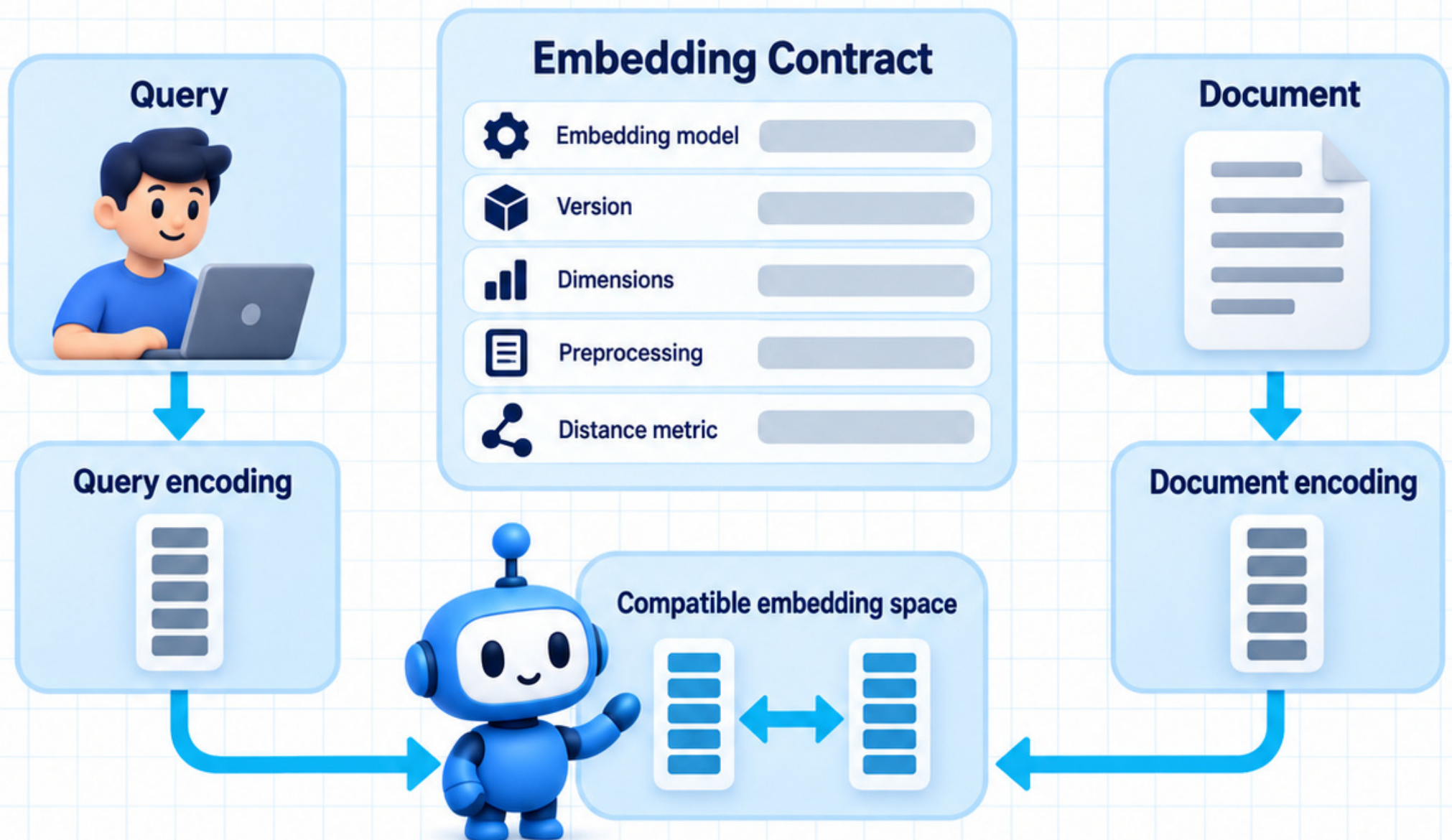
An embedding may connect that wording to a return-policy passage.
Neariness does not prove that a source is correct, current or authorized.



WRITE THE EMBEDDING CONTRACT

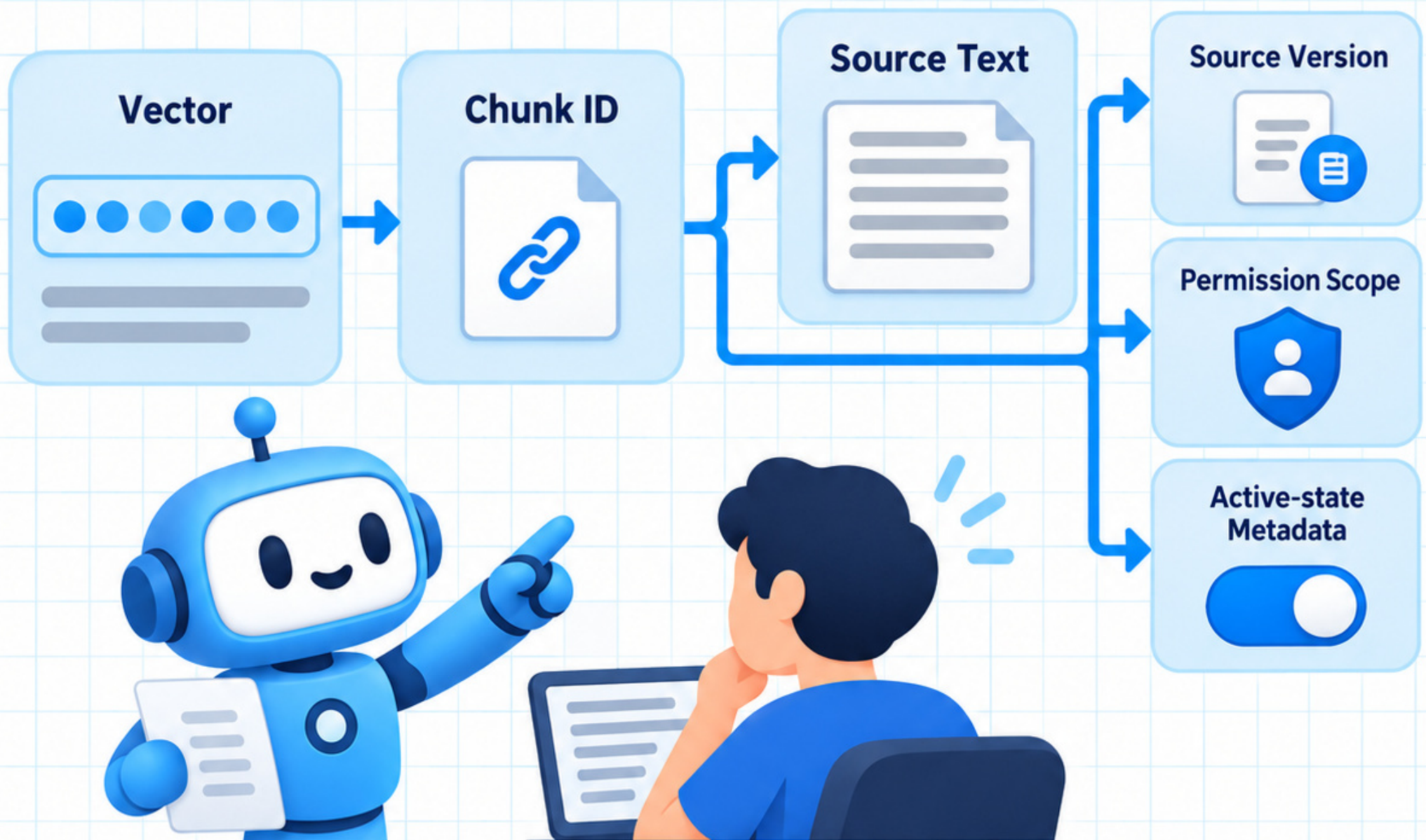
Record the embedding model and version, dimensions, preprocessing and distance metric.

Use the compatible query and document encoding modes.
Equal vector length alone does not make two embedding spaces compatible.



STORE THE EVIDENCE LINK

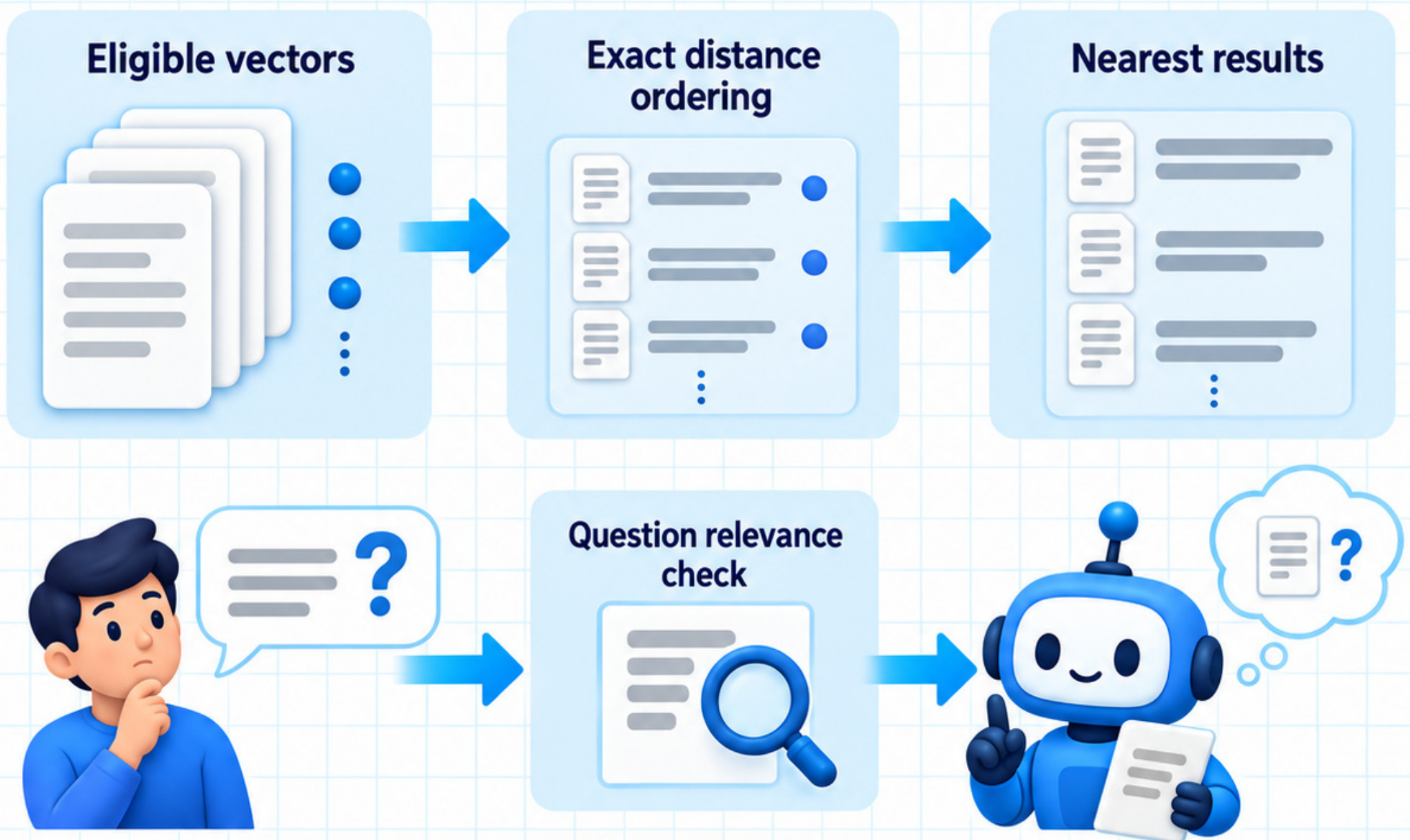
Keep the vector with its chunk ID and embedding version.
Join to source version, permission scope and active-state metadata.
A search result must lead back to the actual supporting text.



BEGIN WITH AN EXACT BASELINE

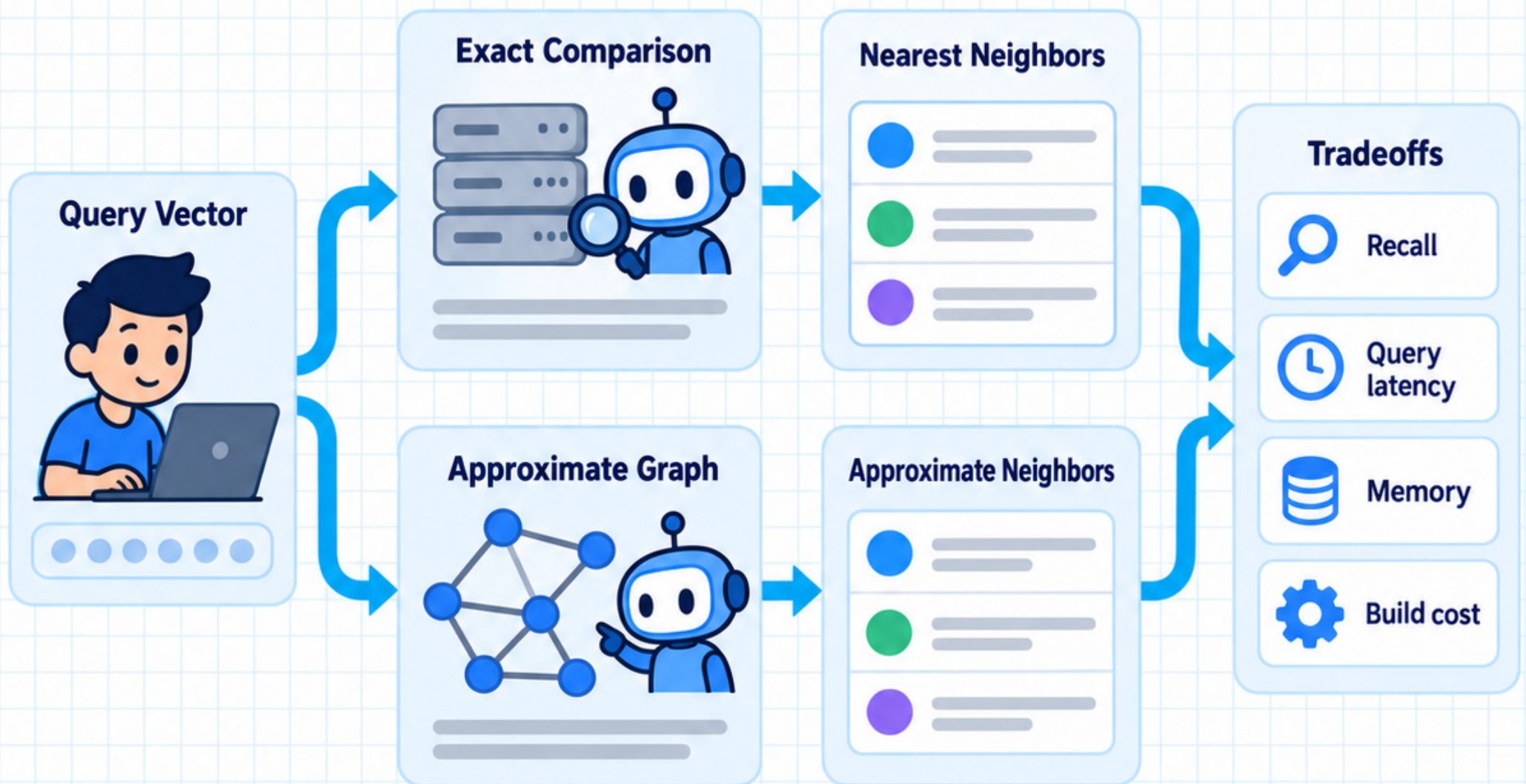
Exact search compares the eligible stored vectors under the chosen distance. Use it as a retrieval baseline for your corpus.

Exact nearest neighbors can still be irrelevant to the user question.



APPROXIMATE SEARCH TRADES WORK

Approximate nearest-neighbor search can avoid comparing every vector. It may miss neighbors that exact search would return. Measure recall, query latency, memory and build cost on your workload.



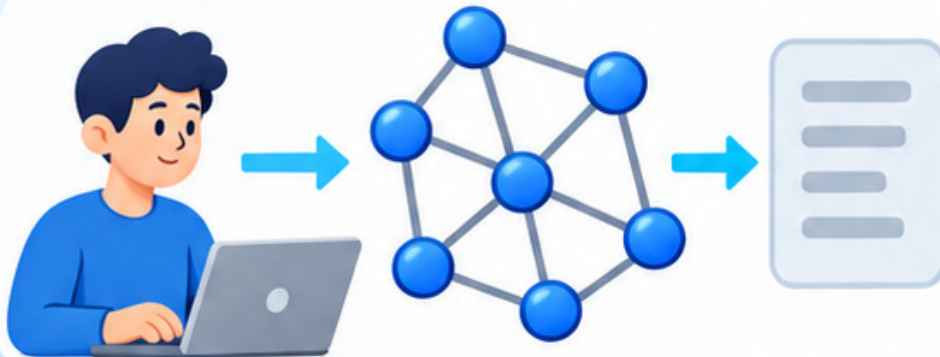
KNOW THE INDEX CHOICES

pgvector supports HNSW and IVFFlat indexes.

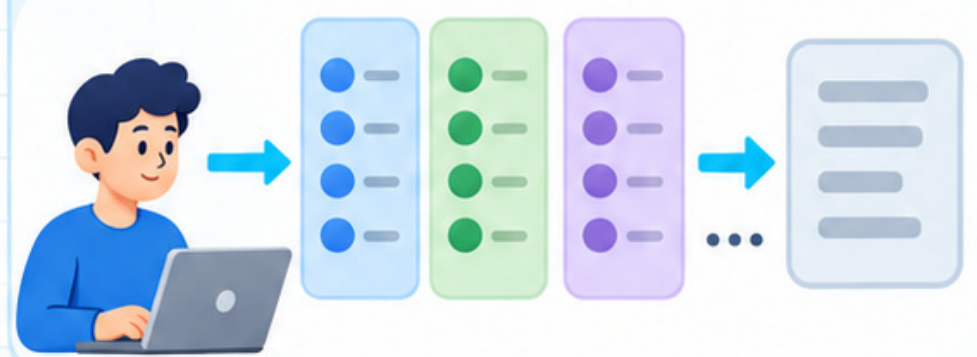
HNSW uses a graph; IVFFlat searches selected vector lists.

Tune and measure the chosen index instead of assuming default settings meet your goal.

HNSW

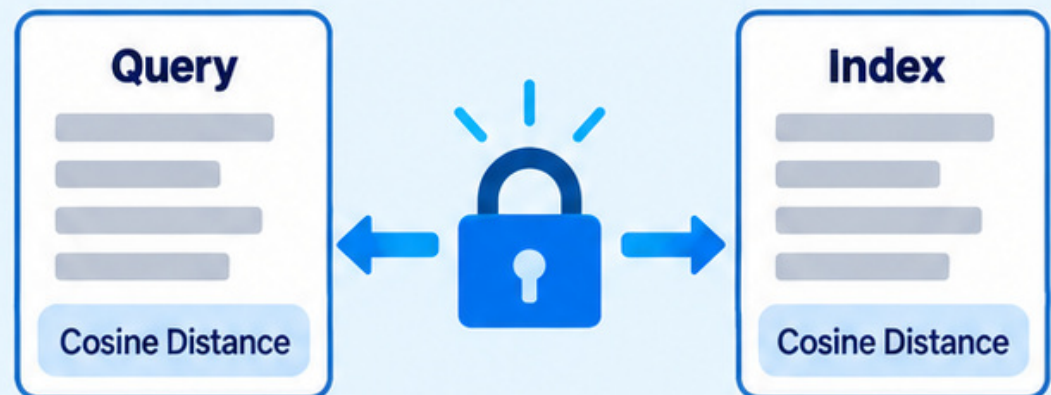


IVFFlat

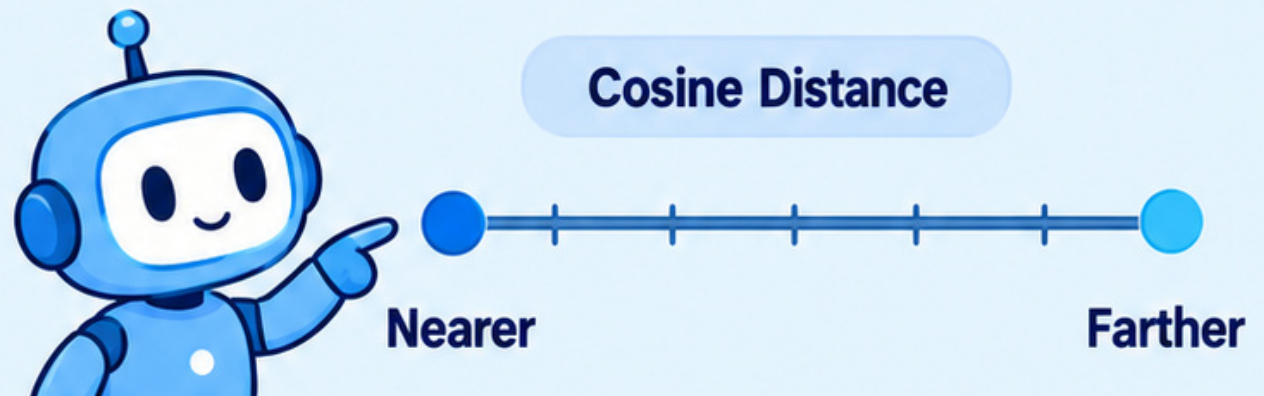


DISTANCE SETTINGS MUST MATCH

The query distance and index operator class must match the design.



For cosine distance, lower values mean closer vectors.



A distance threshold needs task-specific evaluation; it is not a truth score.



FILTERS AFFECT CANDIDATE YIELD



Apply authorization, active-version and validity rules before text reaches model context.



With approximate indexes, filtering can leave too few results.



Measure filtered recall and adjust search effort or index layout.



Approximate
candidate search



Mandatory
eligibility gate



Fewer eligible
results



Context



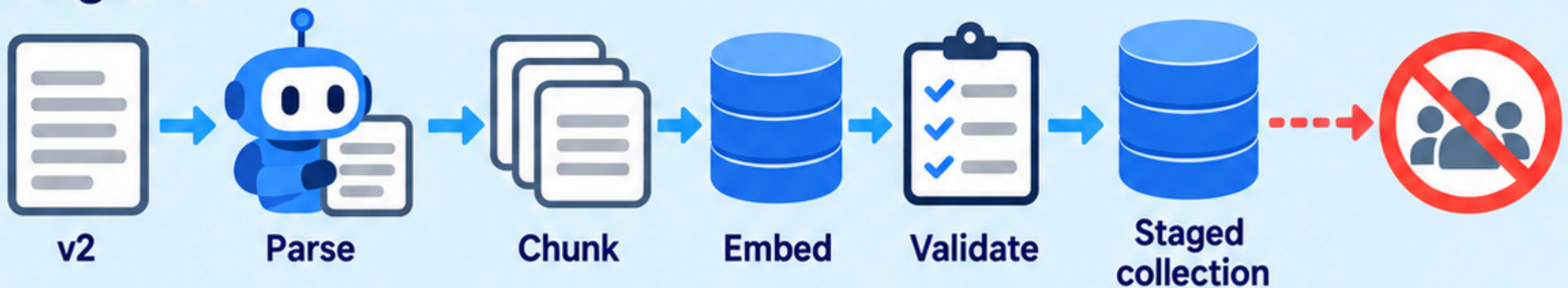
BUILD A STAGED VERSION

Toy policy update: v2 replaces v1.

Parse, chunk, embed and validate v2 in a staged collection.

Serve v1 only while it remains applicable and authorized; otherwise return the defined no-current-evidence response.

Staged v2

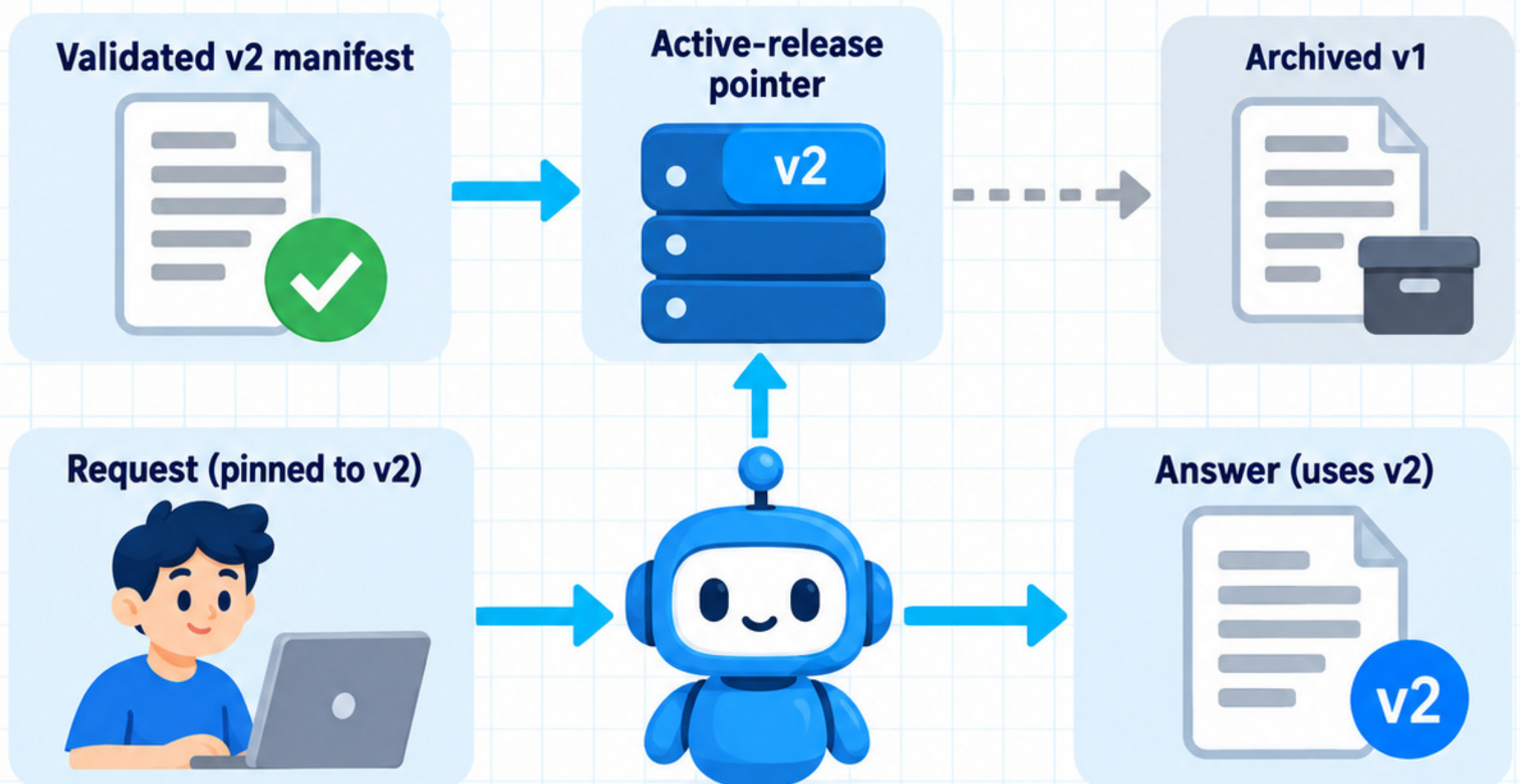


Serving v1



ACTIVATE A COMPLETE RELEASE

- Switch the active release only after its required chunks and metadata are ready.
- Pin each request to one approved release.
- Avoid mixing old and new policy passages in a single answer.

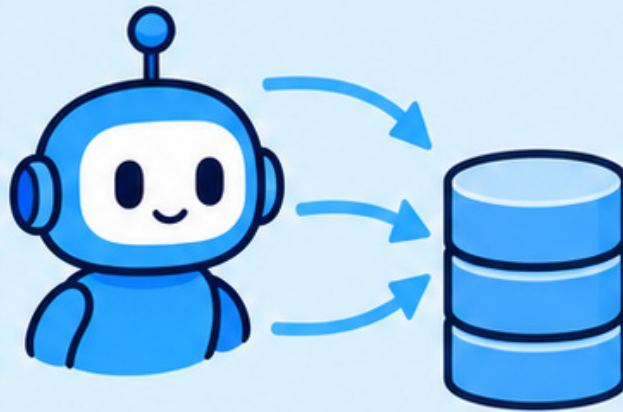


MAKE RETRIES IDEMPOTENT

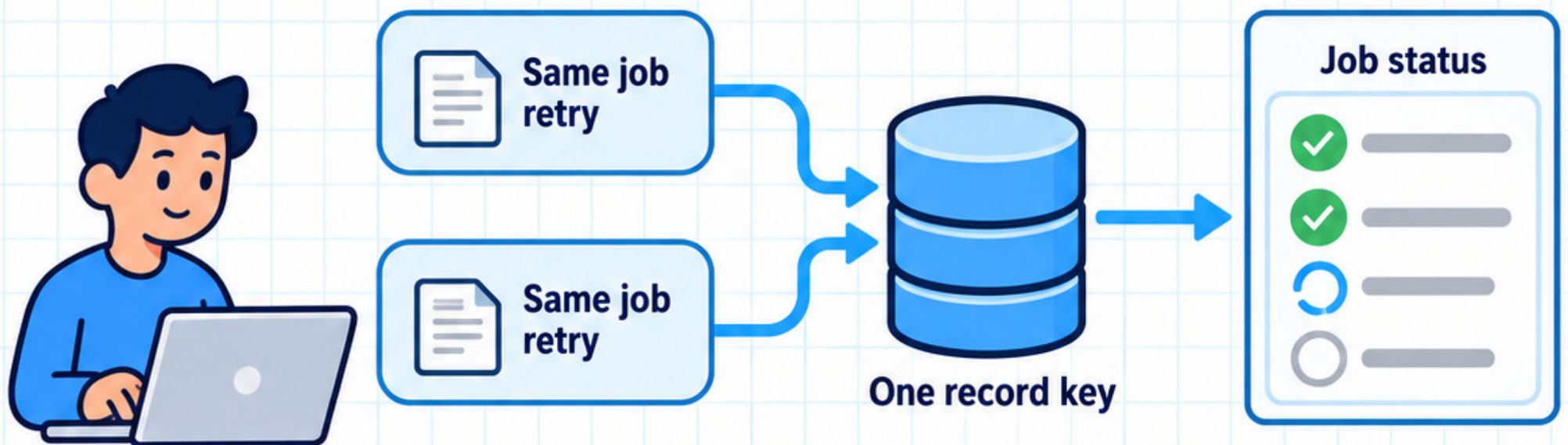
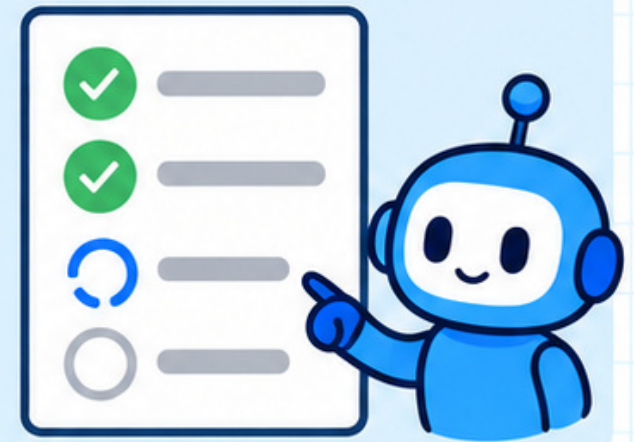
Use stable IDs for the source version, chunk and embedding configuration.



A retried job should update the same intended record.



Record job status so incomplete batches cannot appear complete.

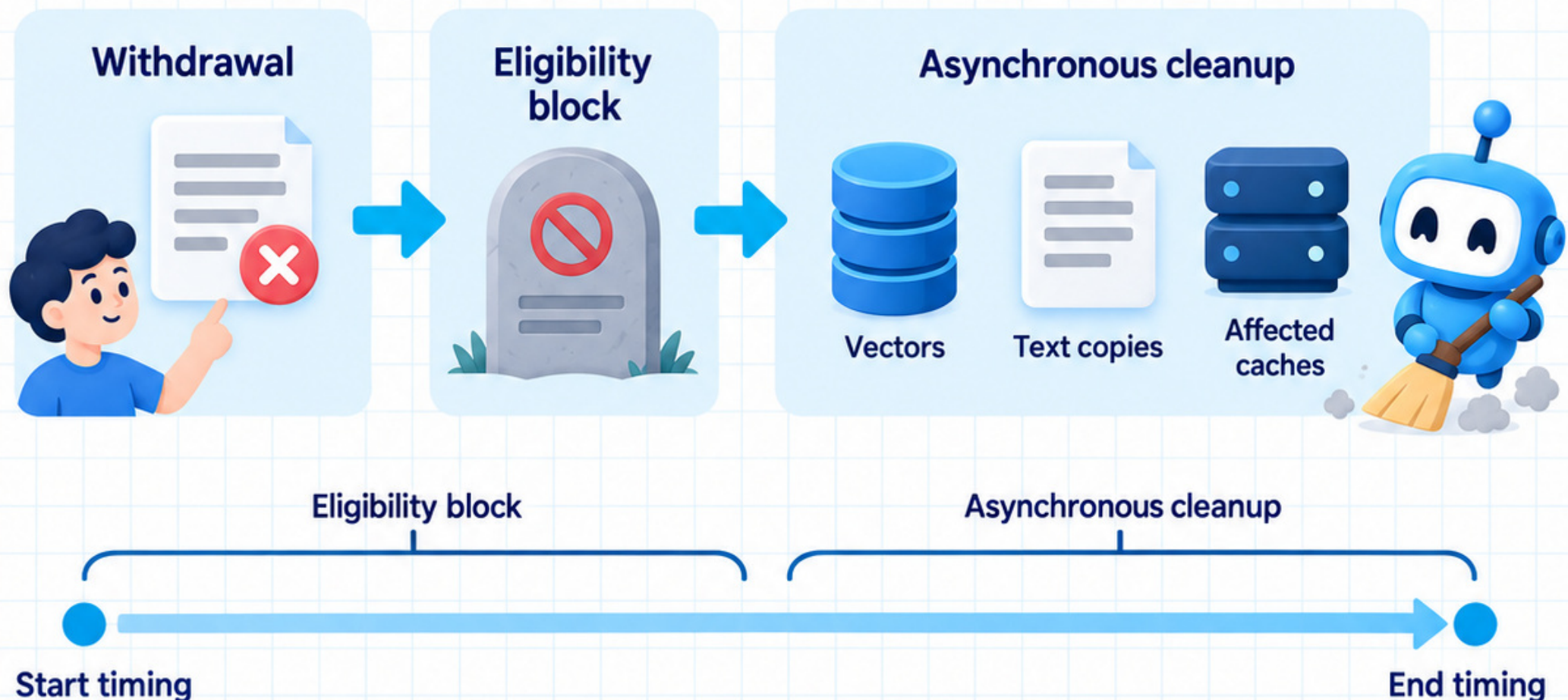


DELETION HAS TWO STAGES

First make deleted or withdrawn content ineligible, for example with a tombstone.

Then remove vectors, text copies and affected caches under the deletion policy.

Measure how long each stage takes.



MIGRATE EMBEDDING SPACES

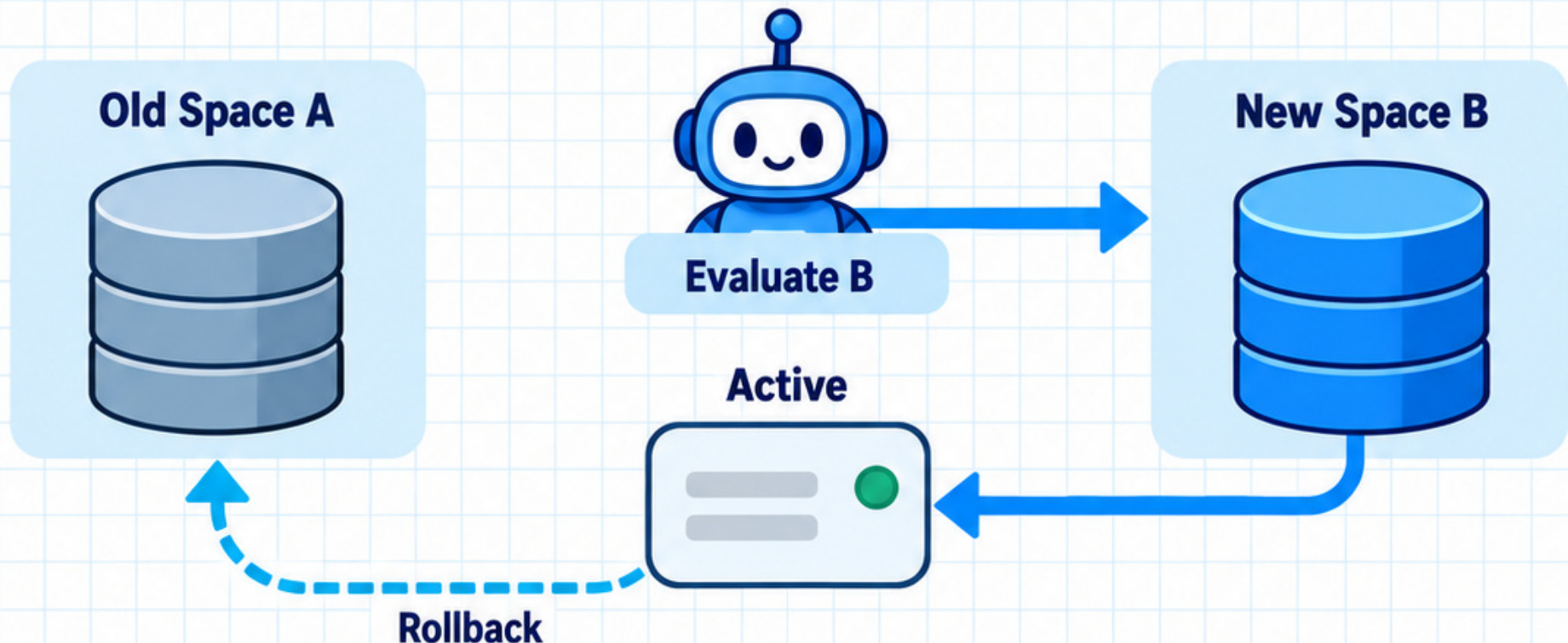
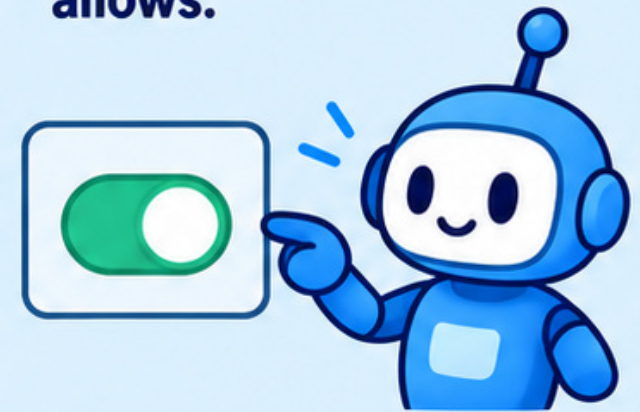
Build a separate index for the new embedding contract.



Evaluate it with compatible query embeddings and the same questions.



Switch a complete release; keep an approved rollback path if policy allows.

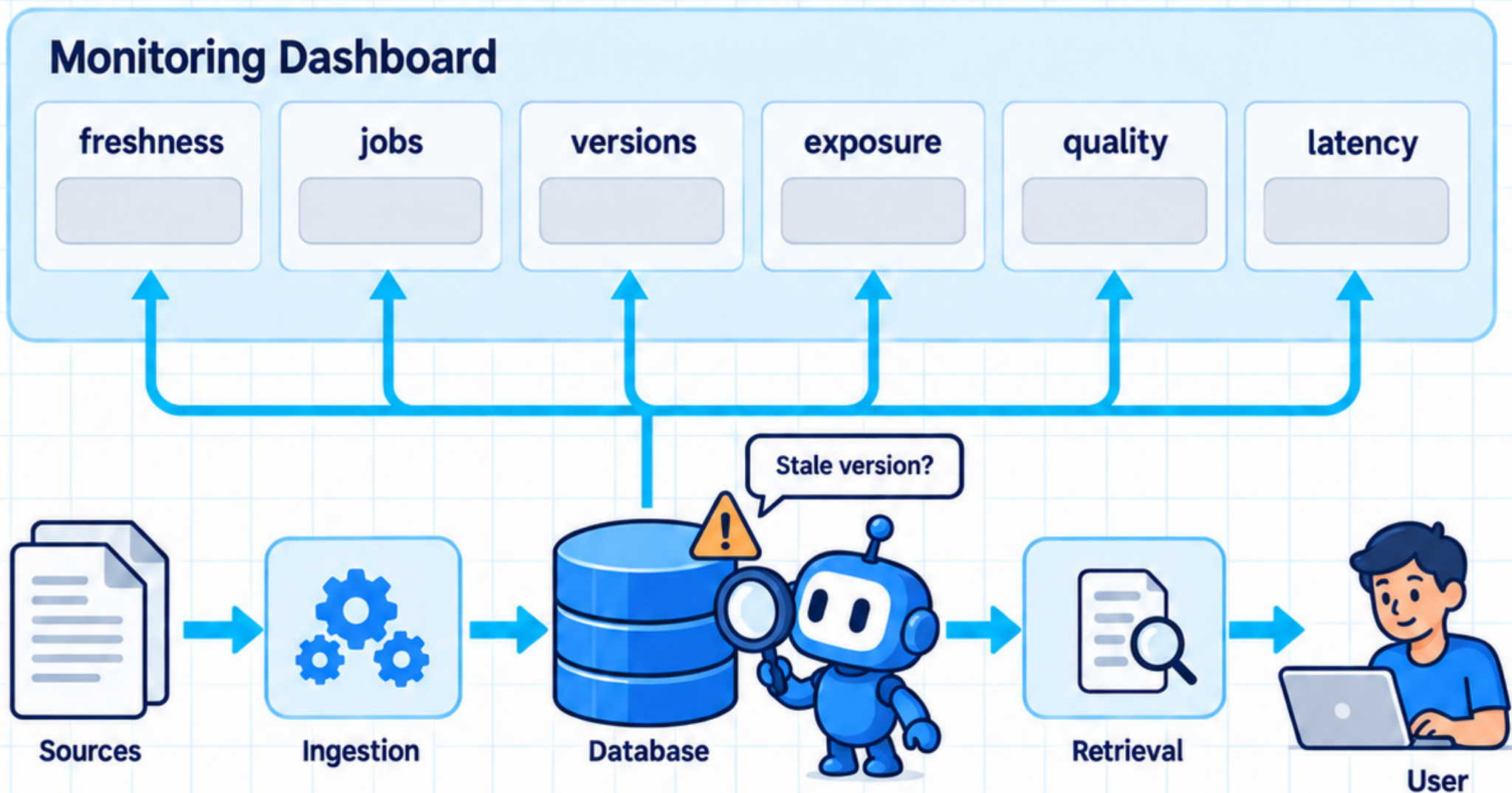


WATCH FRESHNESS AND FAILURES

Track ingestion lag, failed jobs, active source versions and deleted-content exposure.

Also measure filtered retrieval quality and query latency.

A healthy database can still serve outdated evidence.

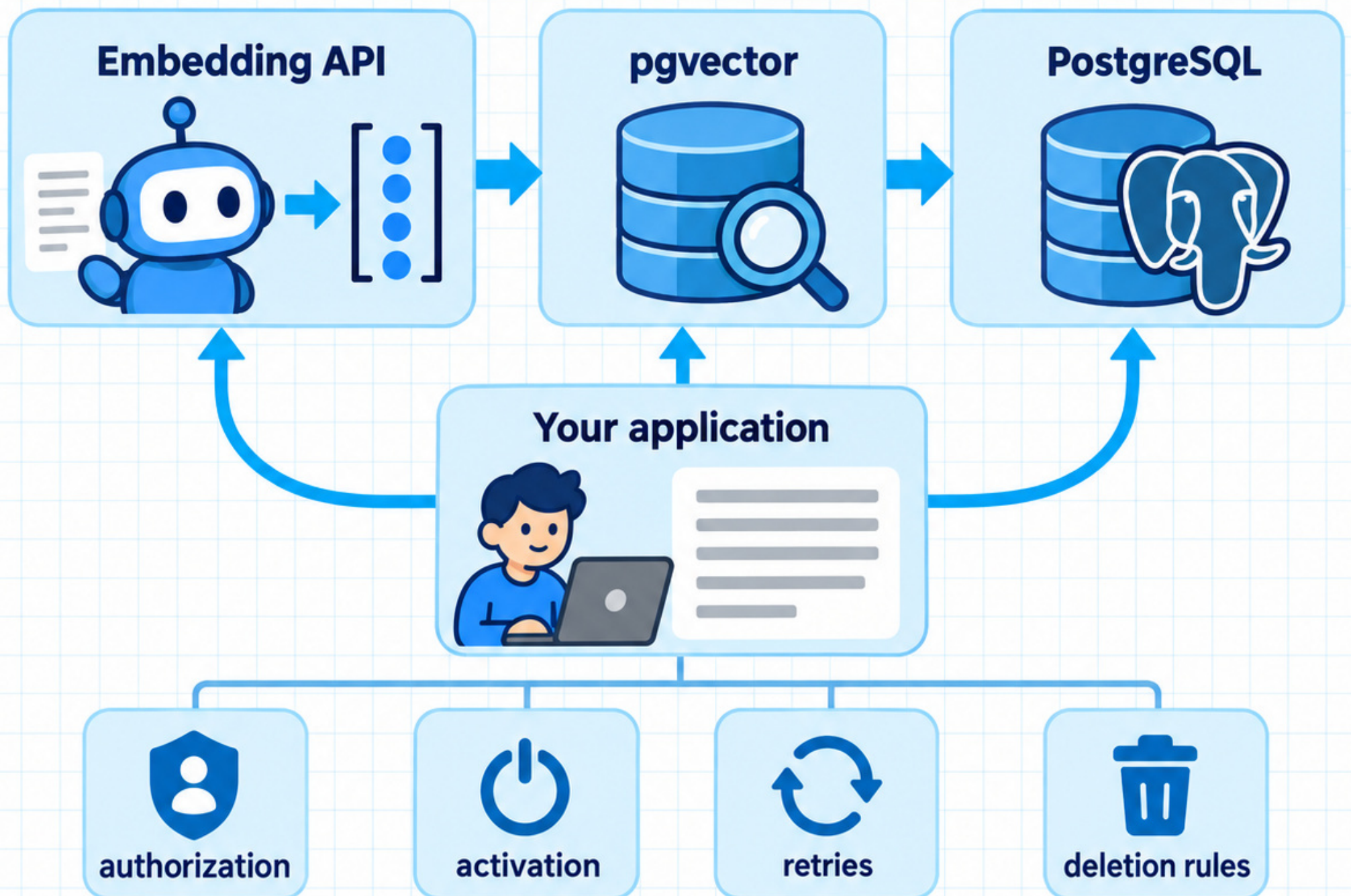


TOOLS AND OWNERSHIP

An embedding API produces vectors. pgvector stores and searches them.

PostgreSQL can track jobs and active releases.

Your application defines authorization, activation, retries and deletion rules.

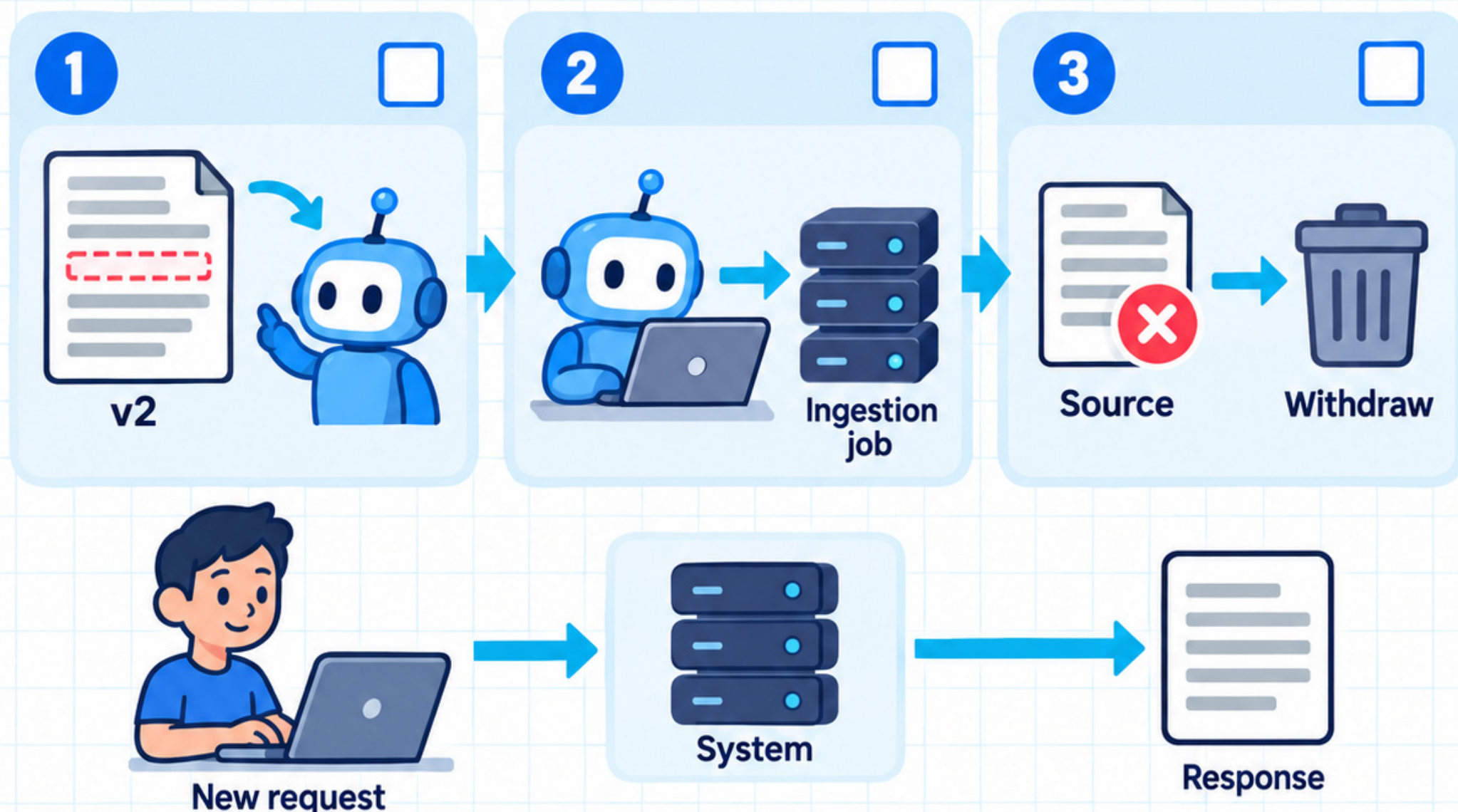


RUN THE UPDATE DRILL

Toy drill: v2 removes a sentence that appears in v1.

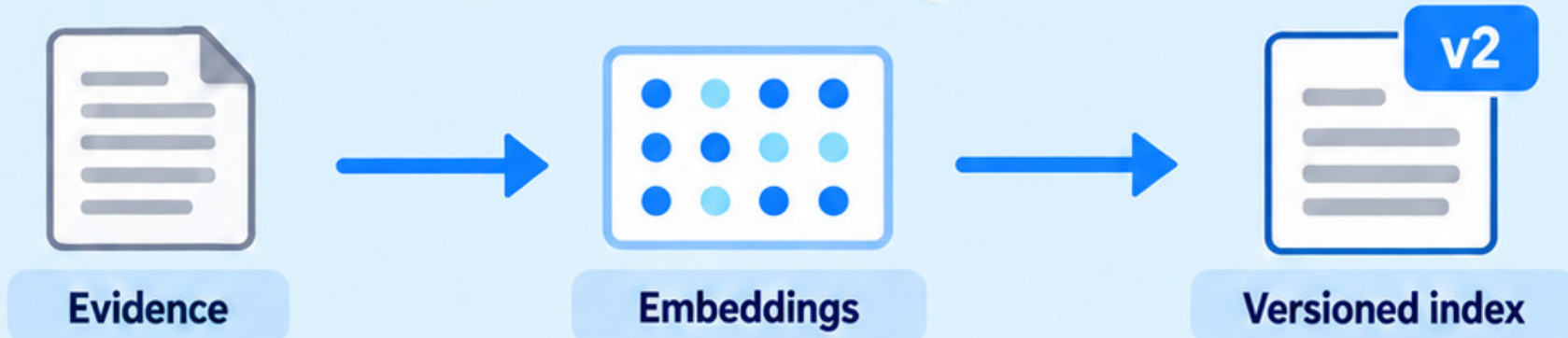
Activate v2, retry its ingestion job, then withdraw the source.

What checks prove that neither duplicates nor withdrawn text reach new requests?

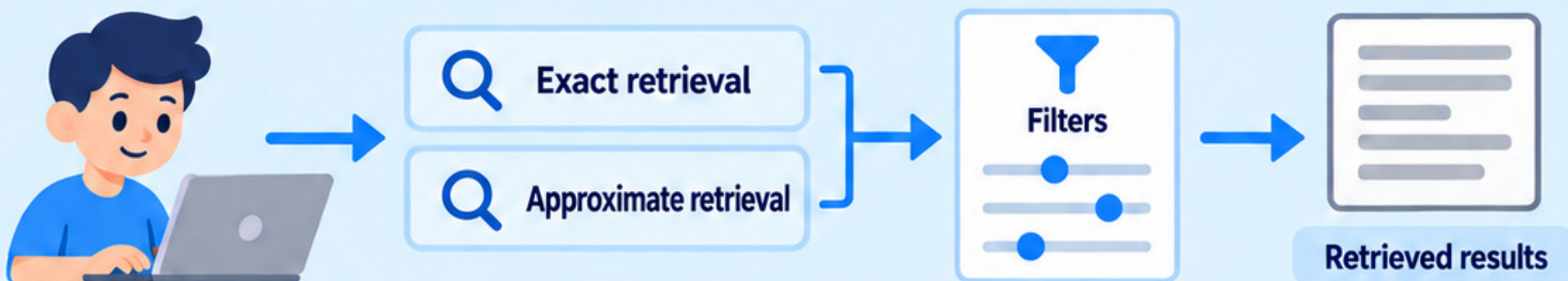


YOUR INDEX CONTRACT

Version the embeddings and evidence together.



Test exact and approximate retrieval with real filters.



Activate complete releases and make withdrawn content ineligible. Next: retrieve and rerank evidence.

