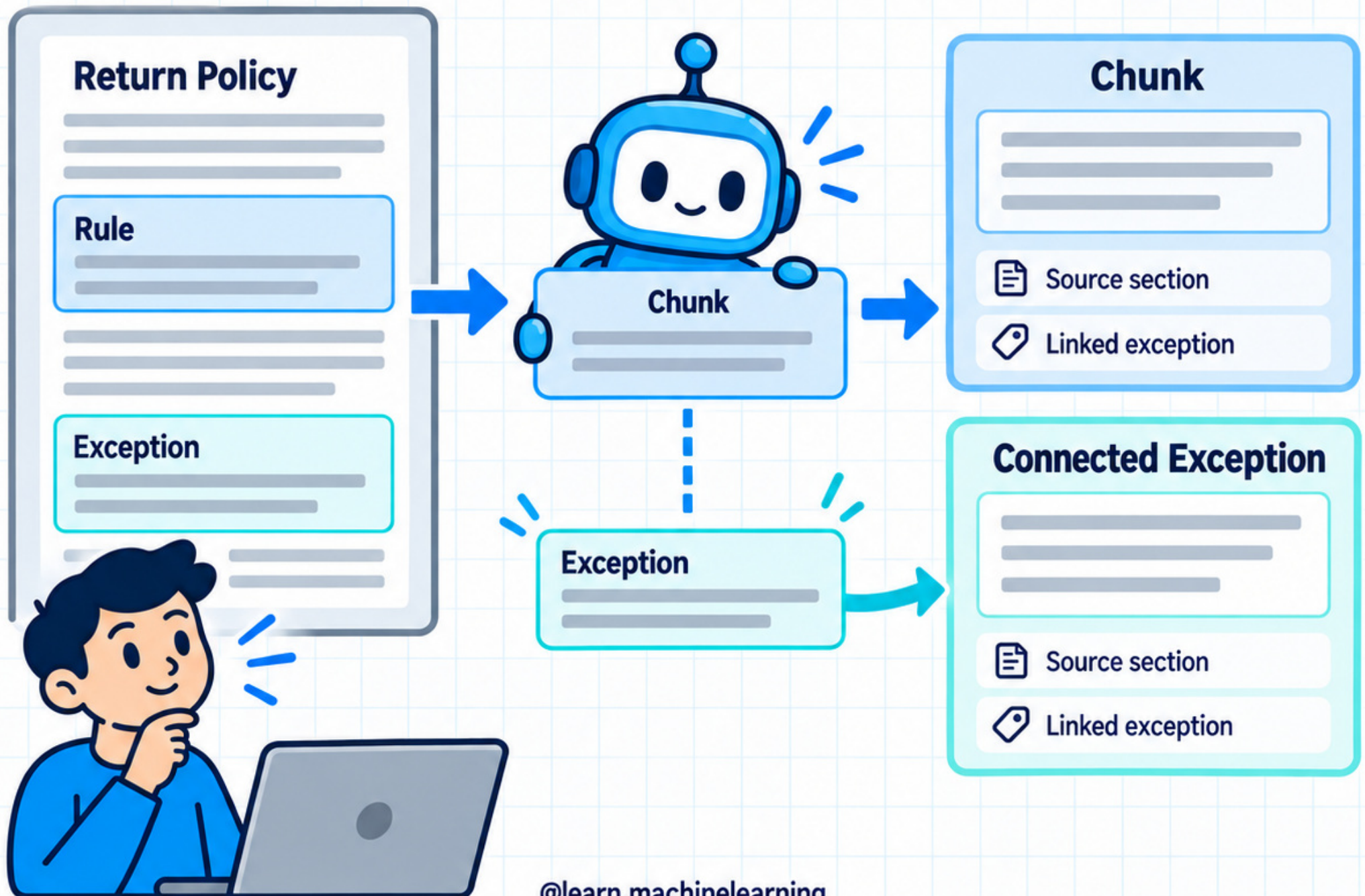
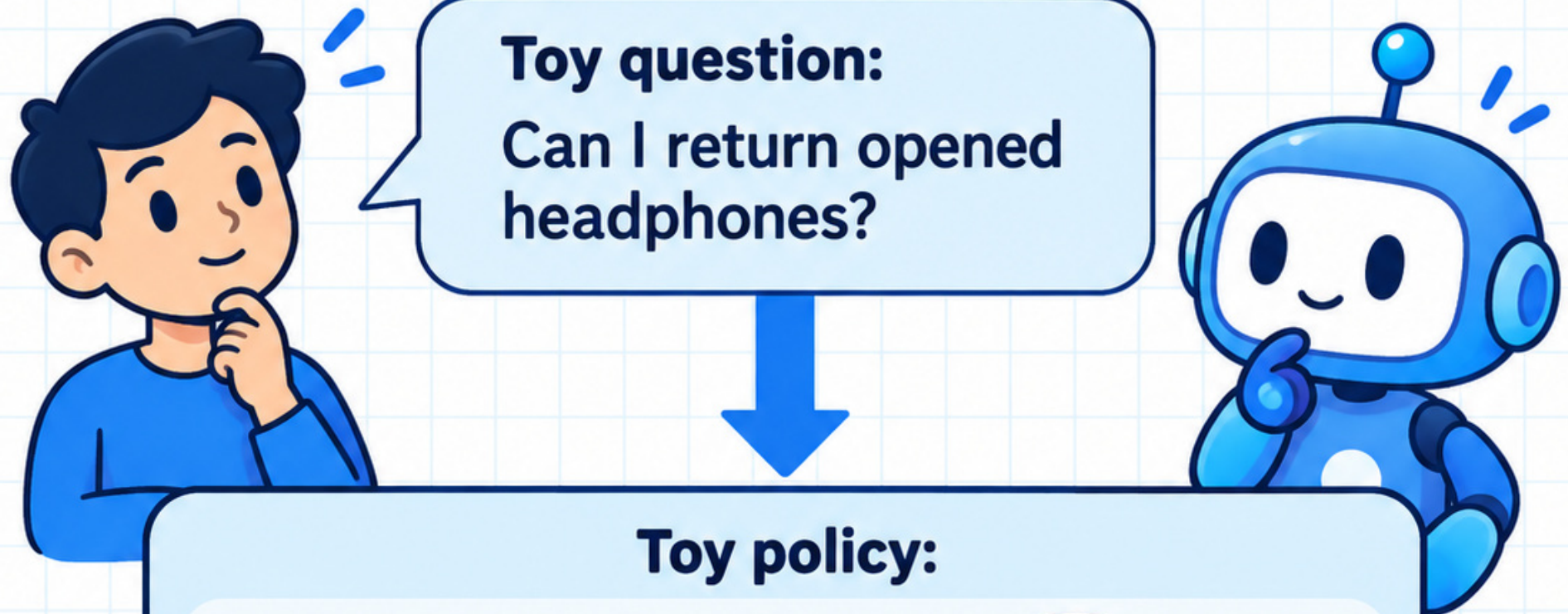


DESIGN CHUNKS AND METADATA

A chunk is a retrievable piece of a source.
The design goal is useful evidence with enough context to interpret it.
Today: split a fictional return policy without losing its exceptions.



START WITH A QUESTION



Toy question:
Can I return opened headphones?

Toy policy:



Unopened items
may be returned
within 30 days.



Opened headphones
are excluded unless
defective.

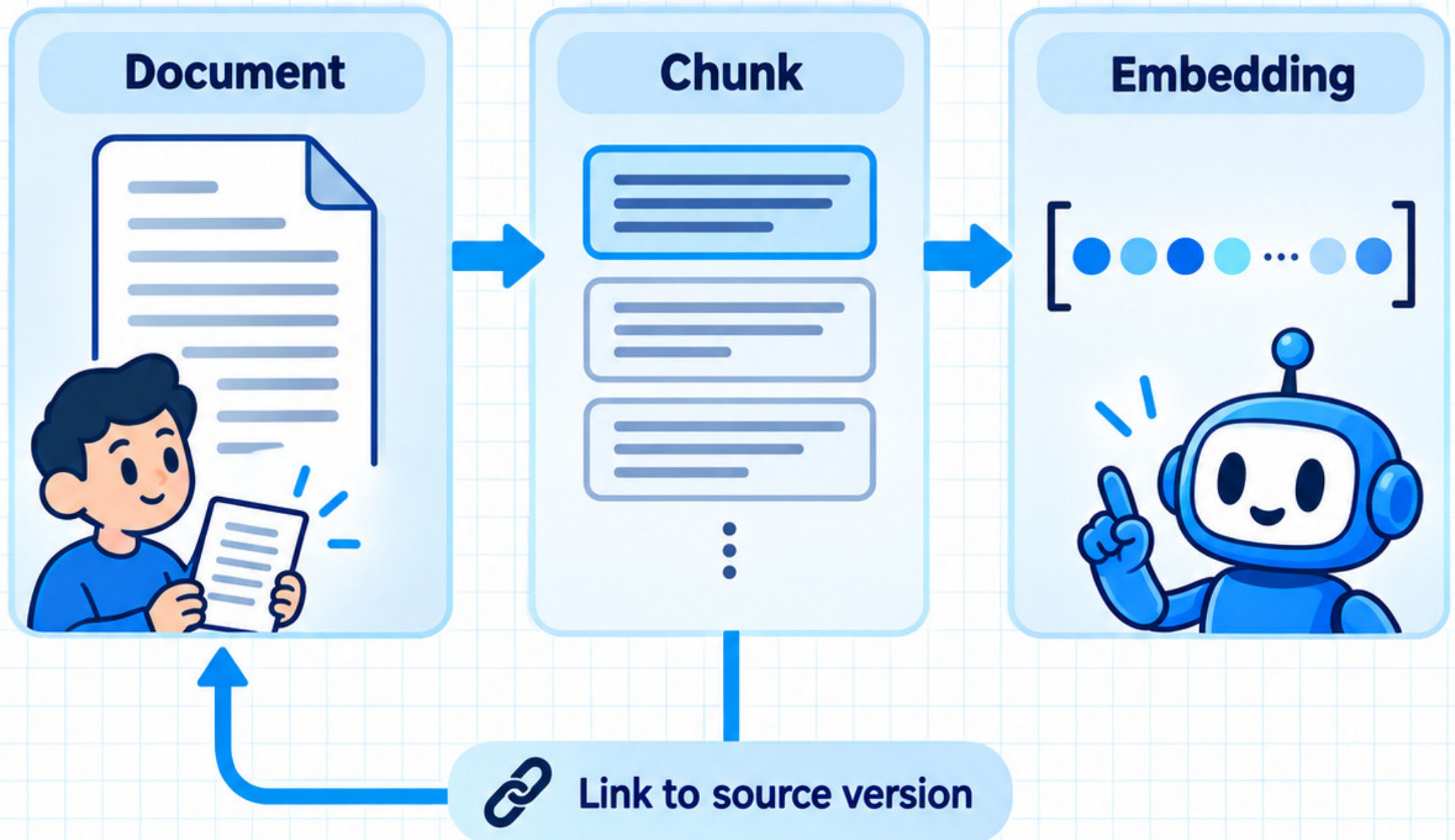
A useful chunk must preserve the exception.

CHUNKS ARE DESIGN UNITS

A document is a source artifact. A **chunk** is a retrieval unit.

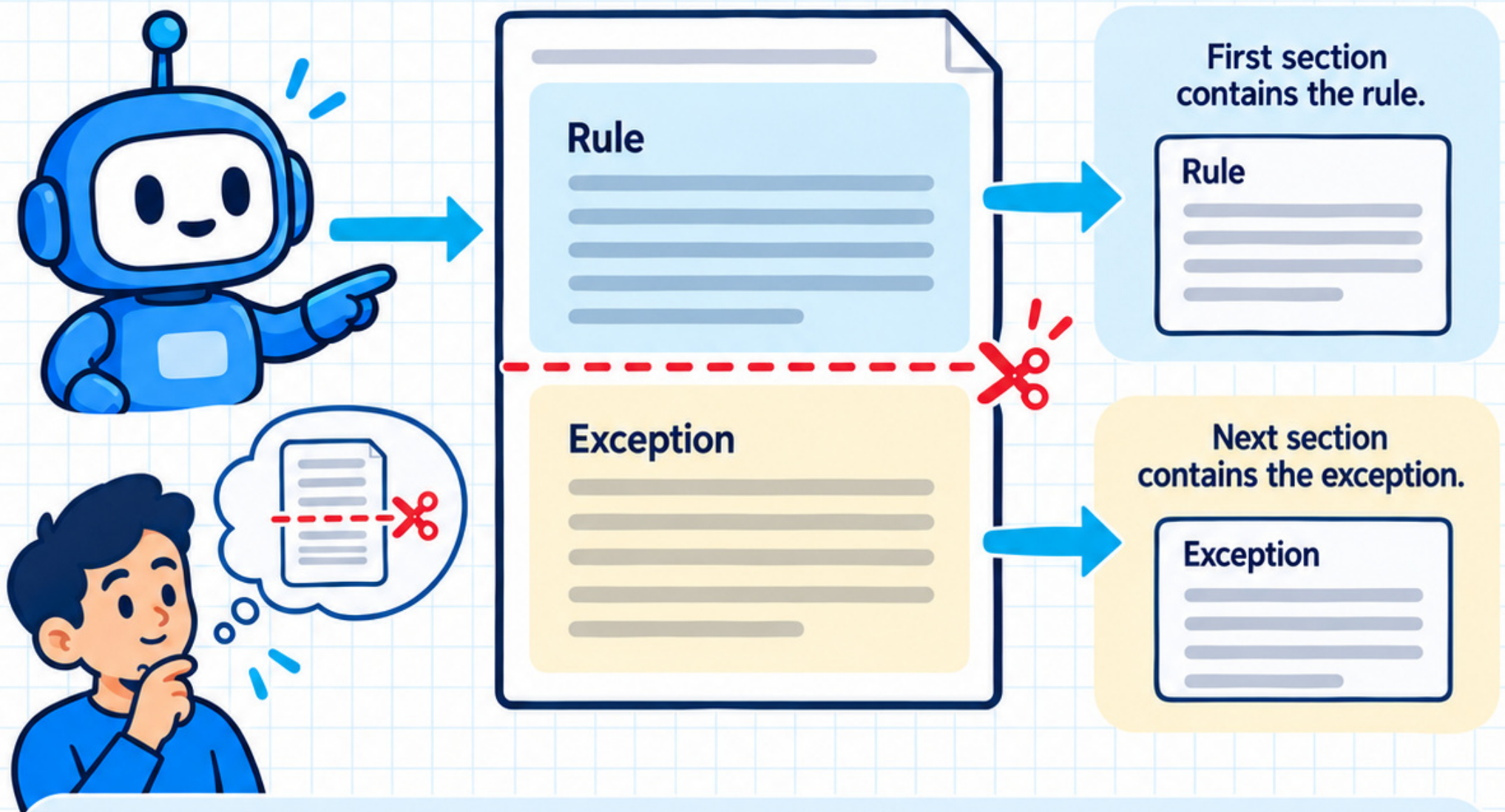
An **embedding** is a numerical representation used for similarity search.

Keep a stable link from every chunk back to its source version.



FIXED-SIZE BOUNDARIES

Split by a measured token or character limit.

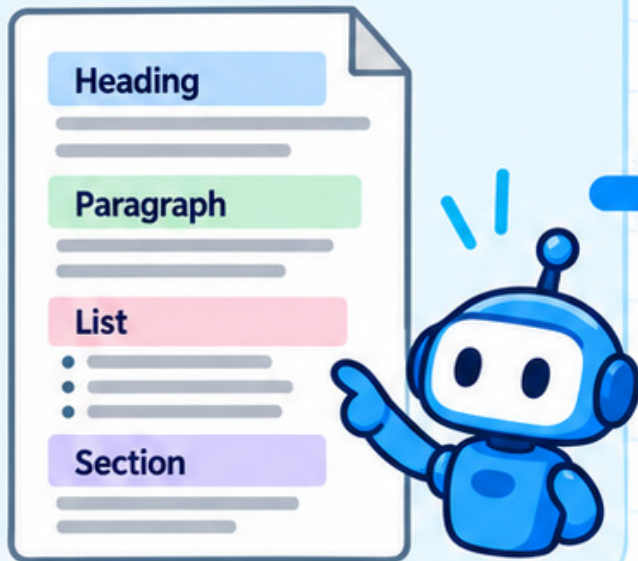


This is simple to implement, but a boundary can separate a rule from its exception.

Specify the unit and test what each boundary cuts.

USE DOCUMENT STRUCTURE

Split around headings, paragraphs, lists or sections.



Keep the section title with its content.



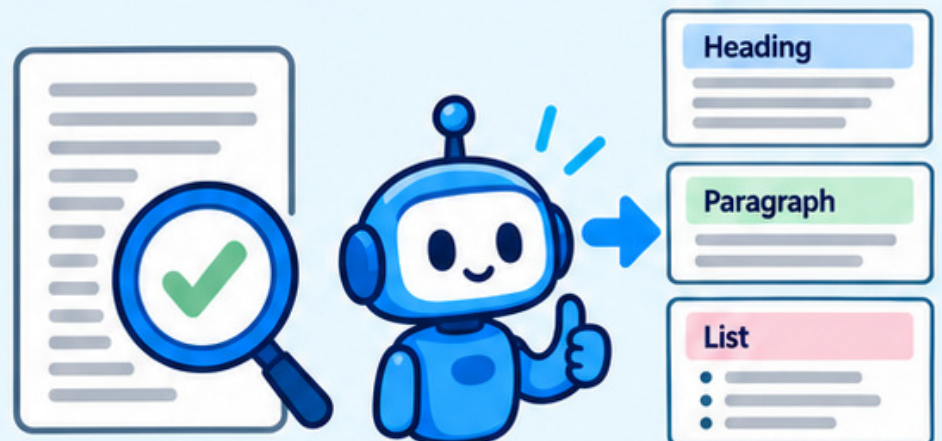
Structure helps only when parsing preserves the actual reading order.



Rule + exception

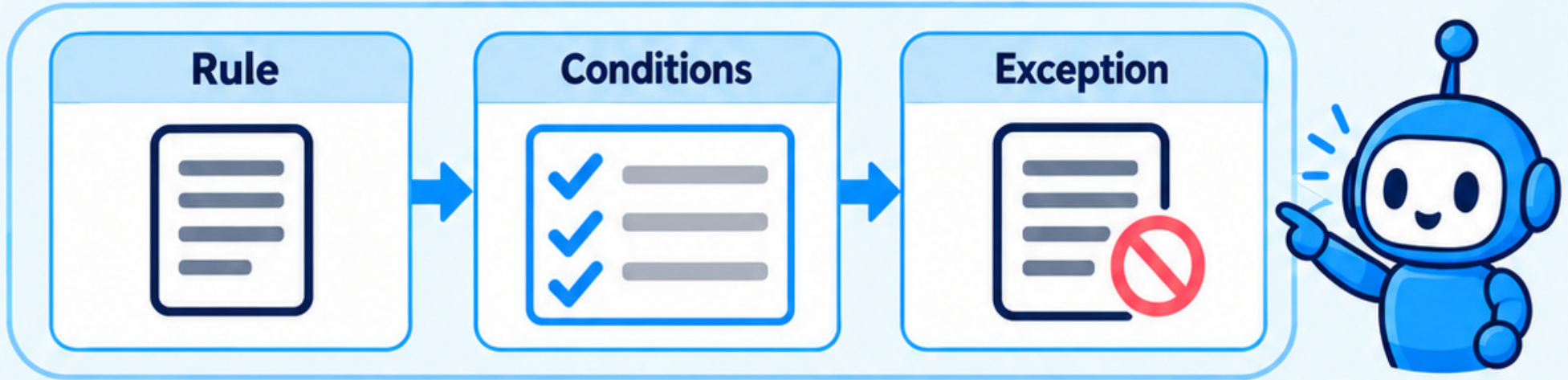


Parse check → Structural split

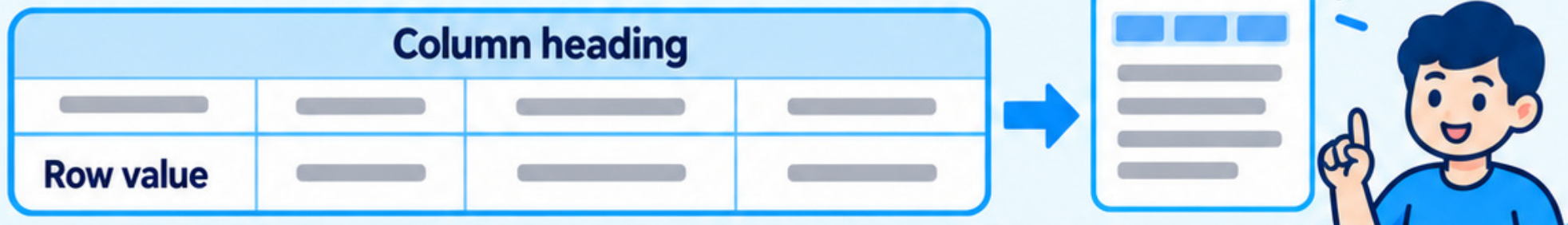


KEEP RELATED FACTS TOGETHER

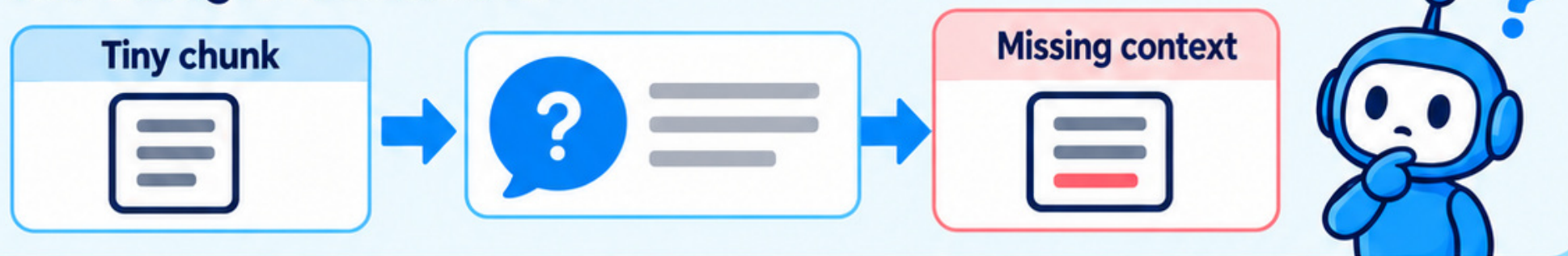
A rule, its conditions and its exceptions form one meaning unit.



A table row may need its column headings.



A tiny chunk can match the question while omitting the fact that changes the answer.

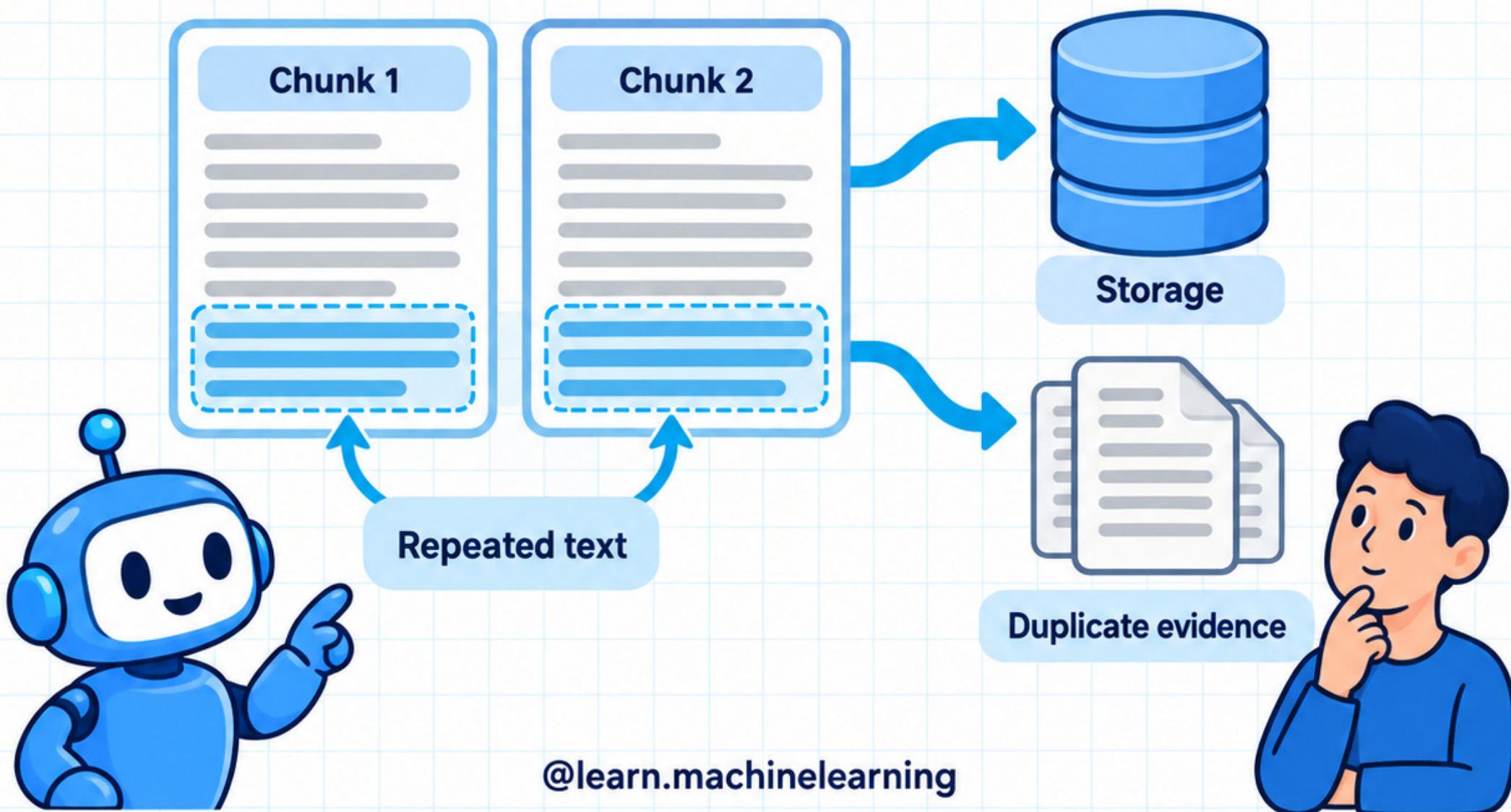


OVERLAP HAS A COST

Overlap repeats text across neighboring chunks.

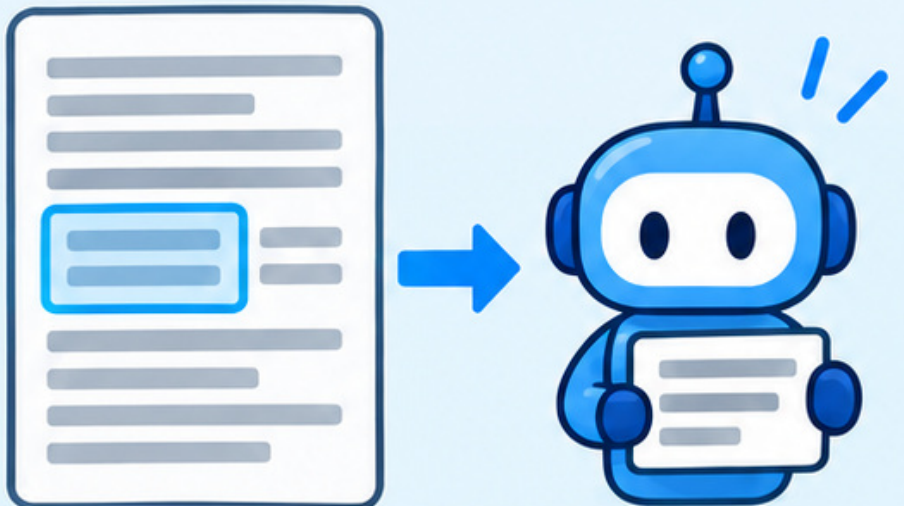
It can protect meaning near a boundary, but increases storage and duplicate evidence.

Measure whether overlap helps your actual questions.



SMALL VERSUS LARGE

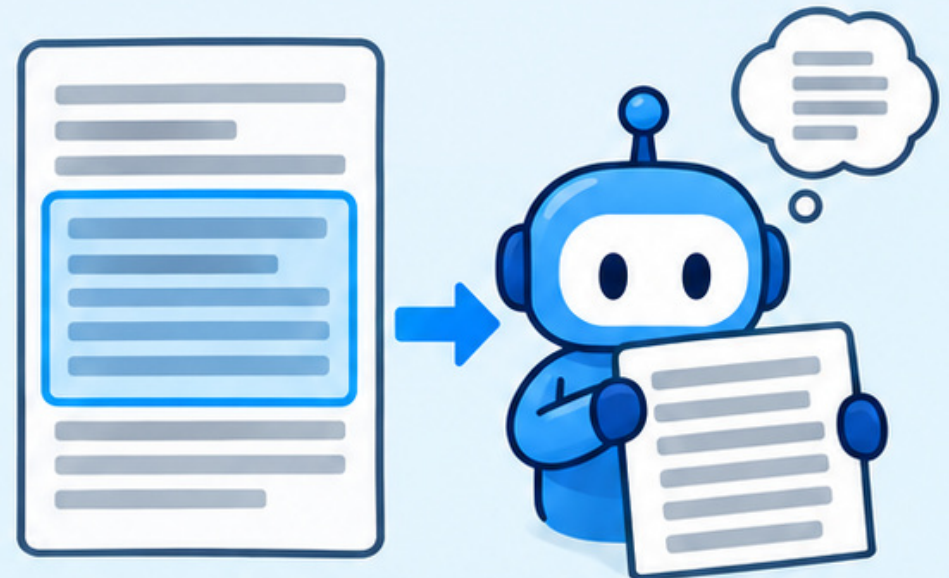
Small chunks can retrieve a focused passage.



Small chunk

Focused passage

Large chunks can retain context but add unrelated text.

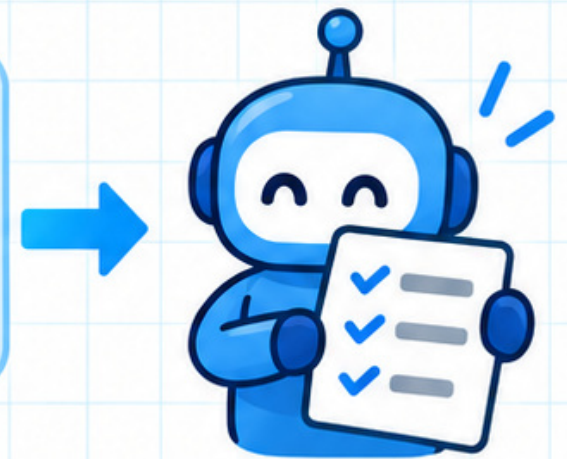


Large chunk

Context with
unrelated text



Choose boundaries using
retrieval and answer checks,
not a universal chunk size.

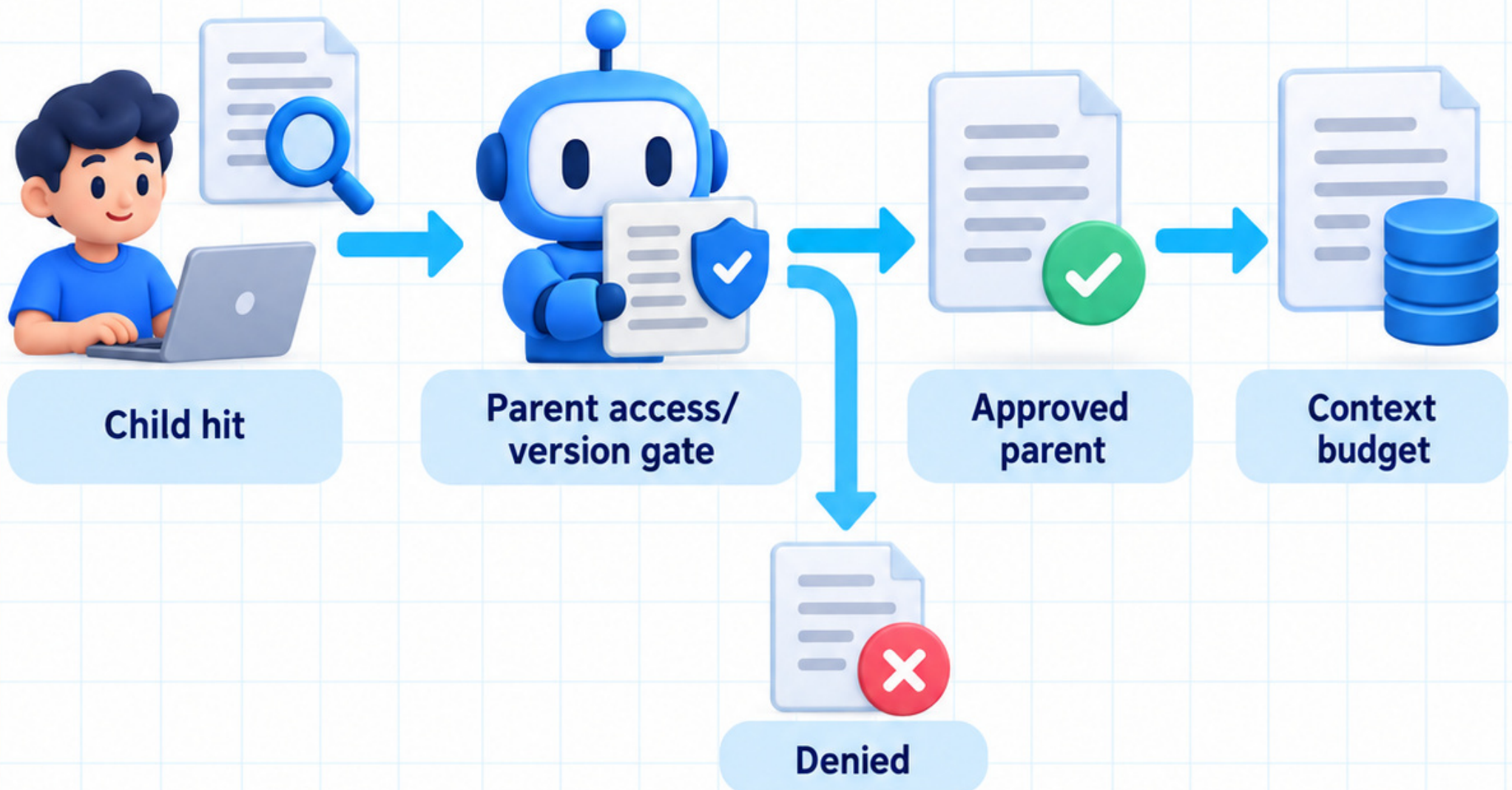


RETRIEVE CHILD, READ PARENT

Search with a focused child chunk, then fetch its parent section when more context is needed.

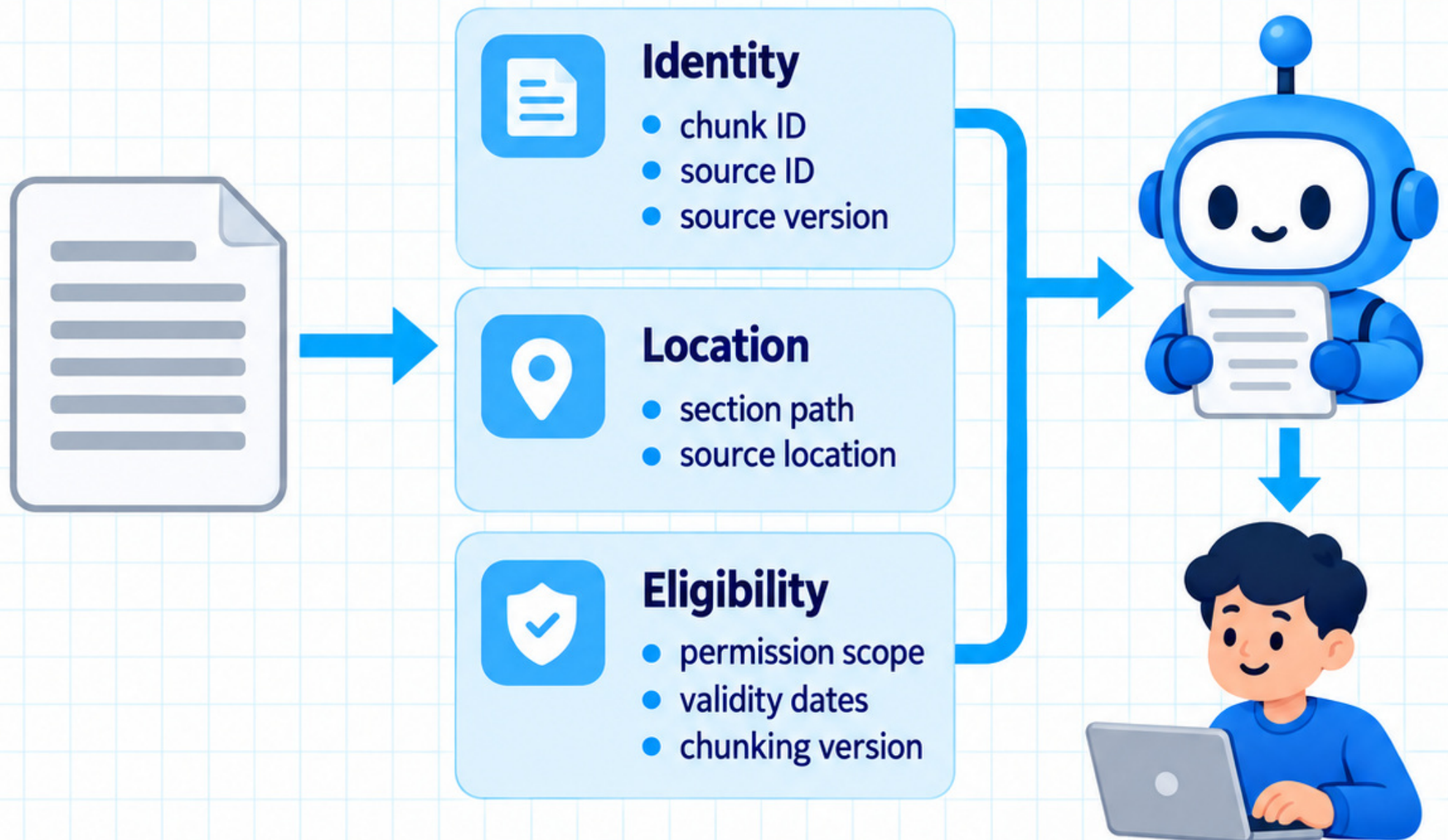
Recheck parent permissions and version before expansion.

Count the expanded text against the context budget.



METADATA MAKES EVIDENCE USABLE

Store chunk ID, source ID, source version, section path and source location.
Also store permission scope, validity dates and chunking version.
The text alone cannot tell the application which policy applies.

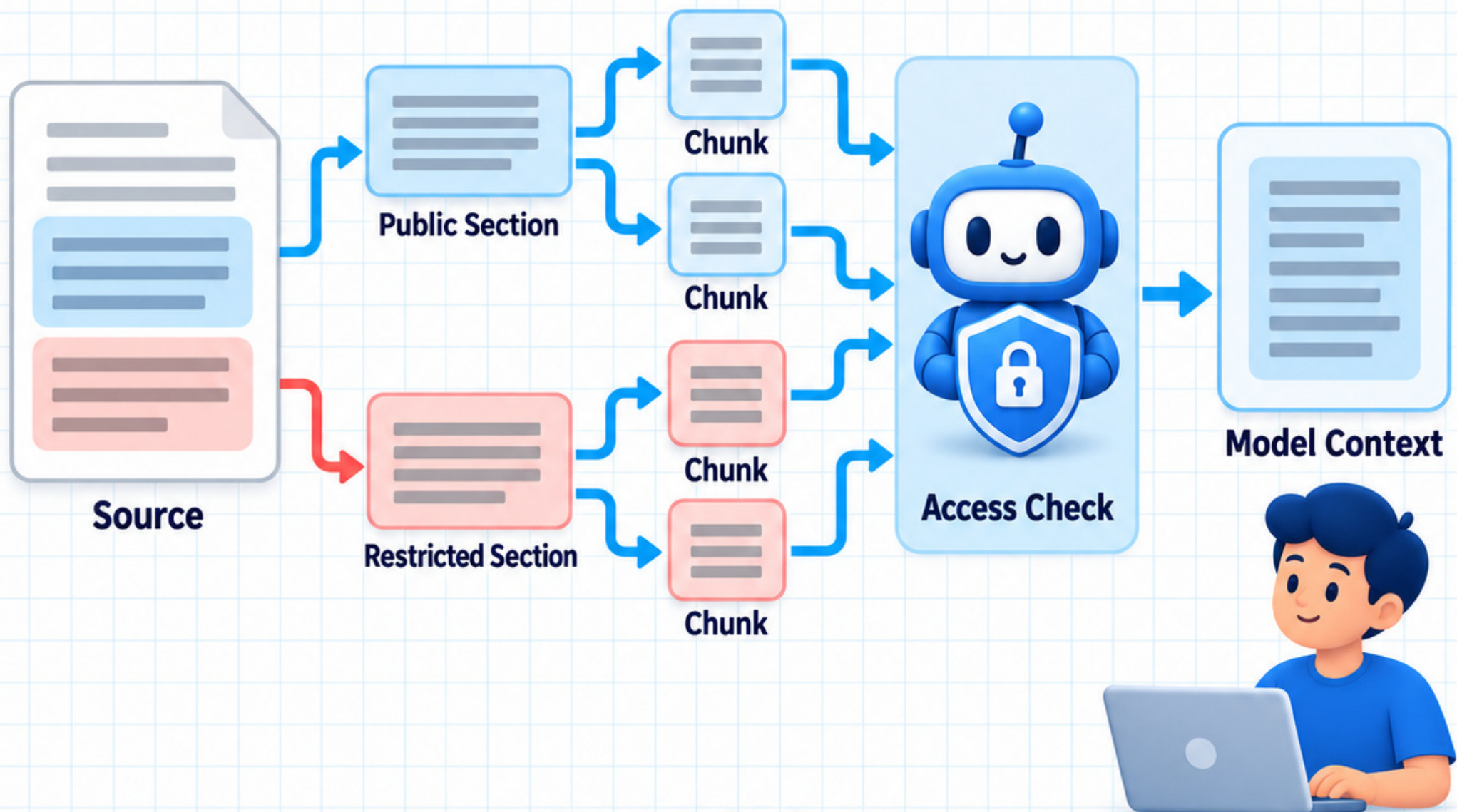


CARRY ACCESS RULES FORWARD

Chunks inherit the applicable source permissions.

If a section is more restricted, preserve that narrower scope.

Authorize every candidate and expanded parent before any text enters model context.

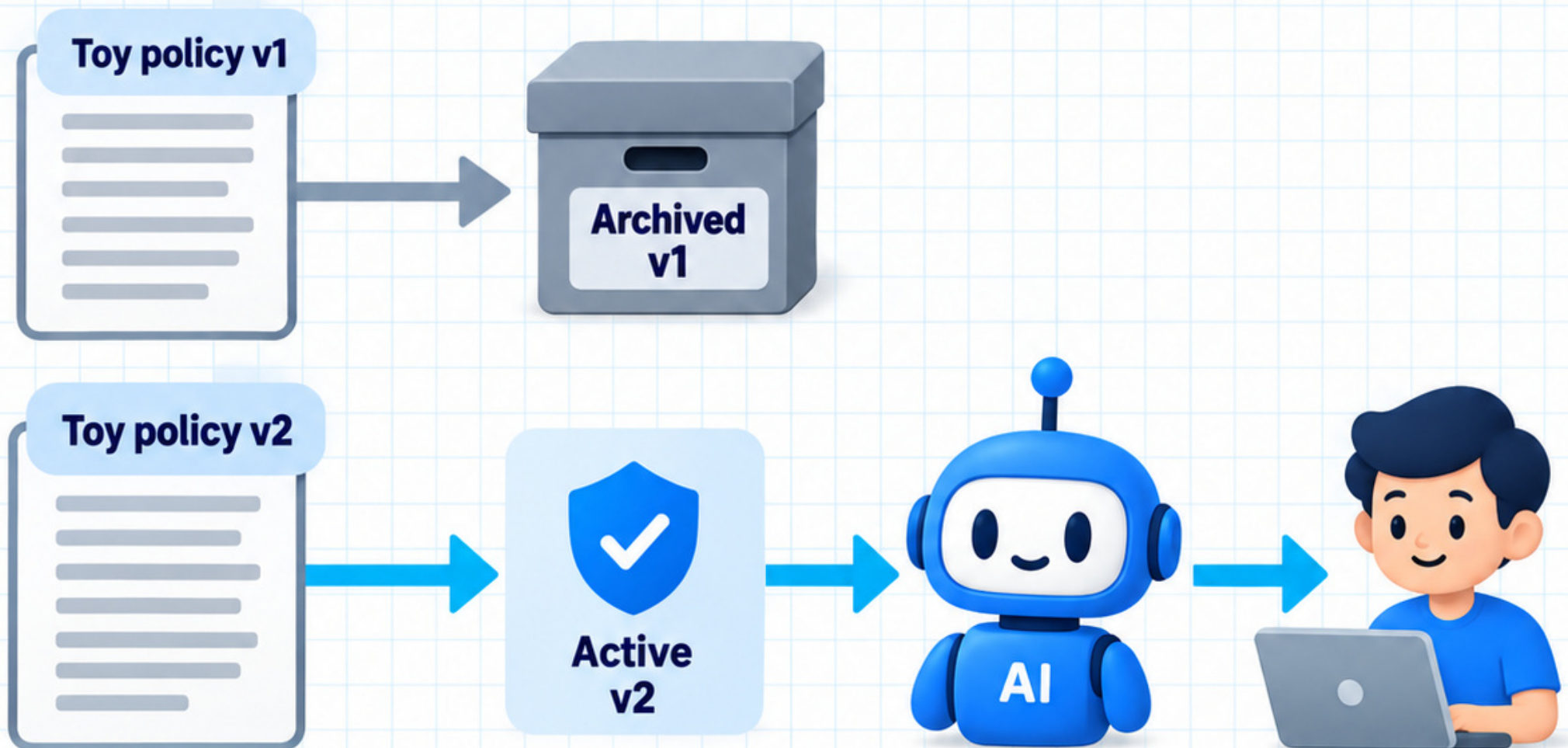


KEEP VERSIONS SEPARATE

Toy policy v1 and v2 are different evidence versions.

Do not merge their paragraphs into one chunk.

Use the active, applicable version for the request and retain provenance for review.

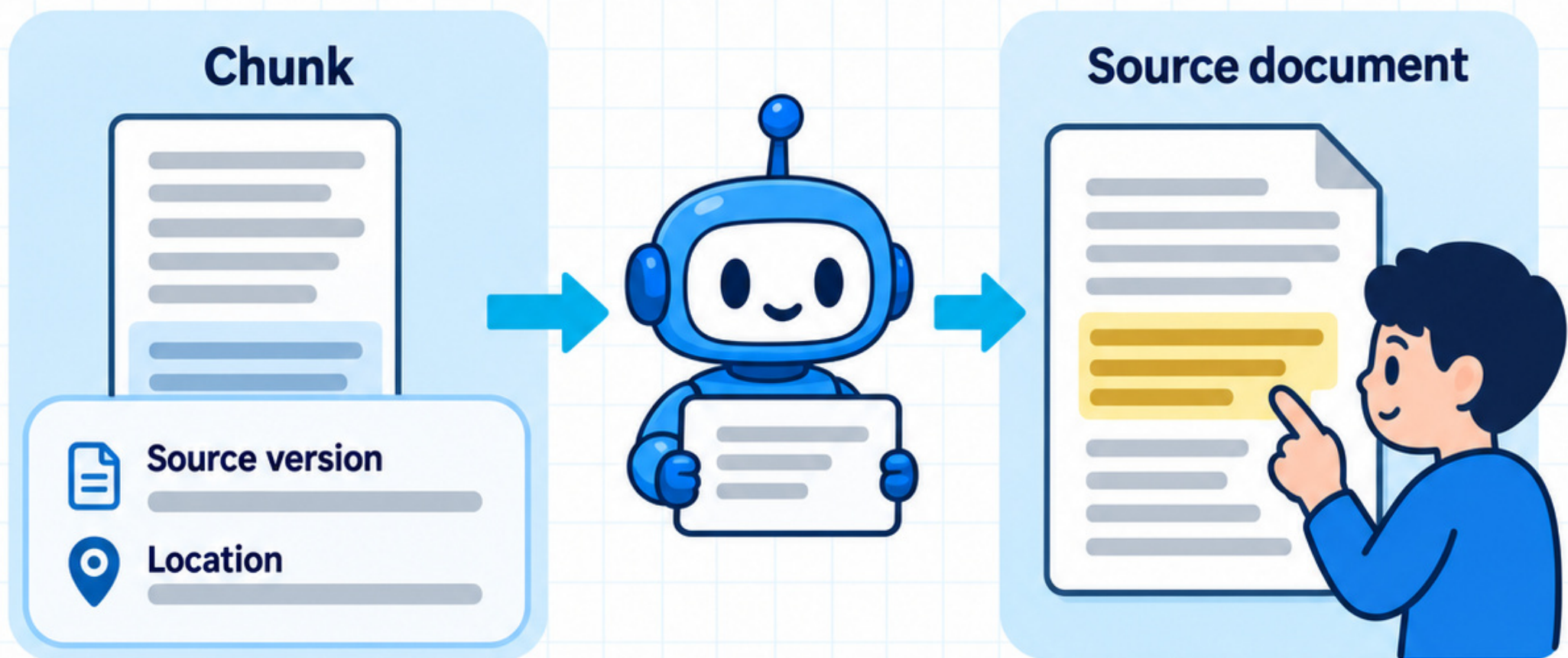


OFFSETS NEED A REFERENCE

A page number, section path or text offset must identify a location in a specific source version.

Parsed-text offsets may differ from PDF coordinates.

Store enough information to reopen the supporting passage.

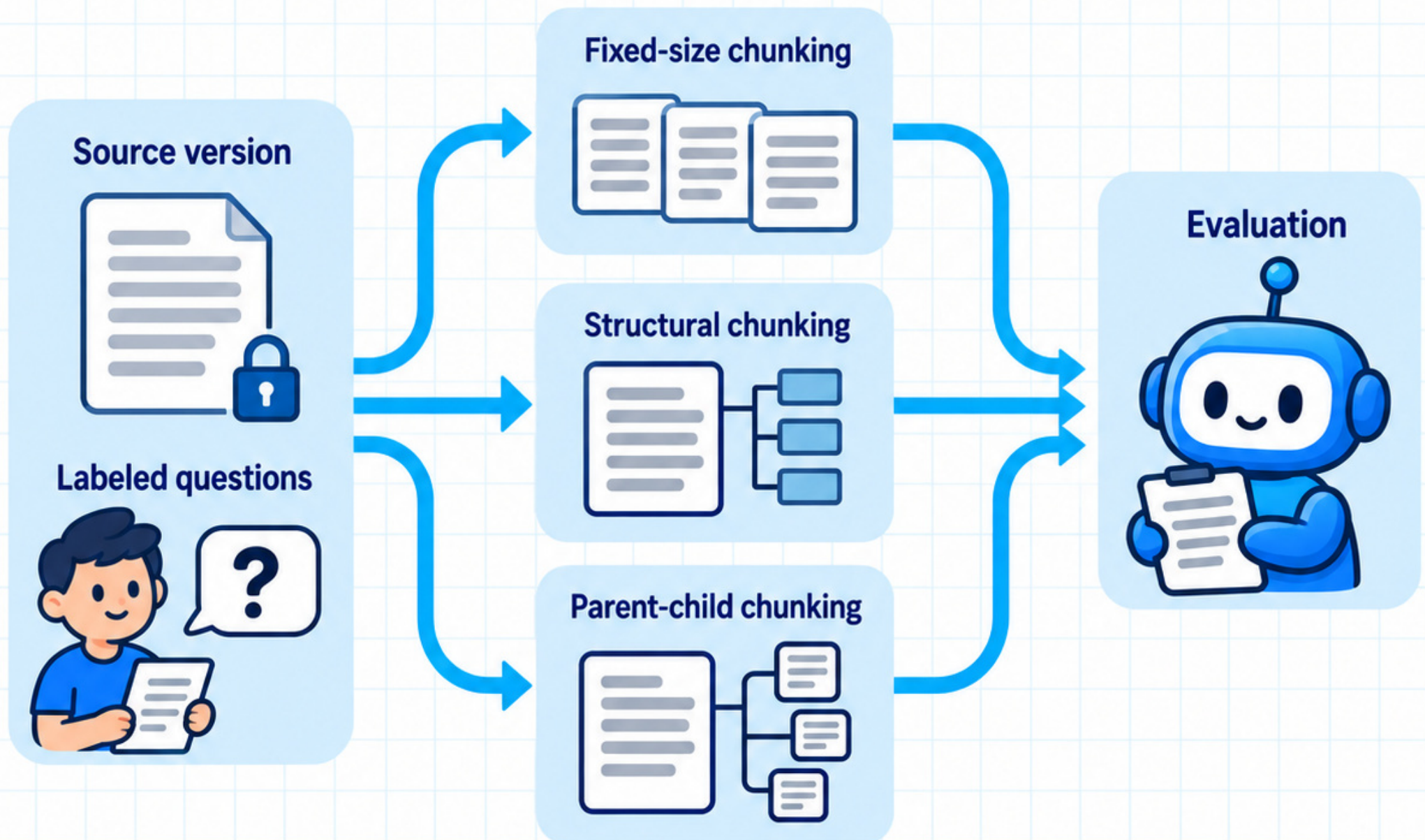


COMPARE THREE VARIANTS

Use the same source version and the same labeled questions.

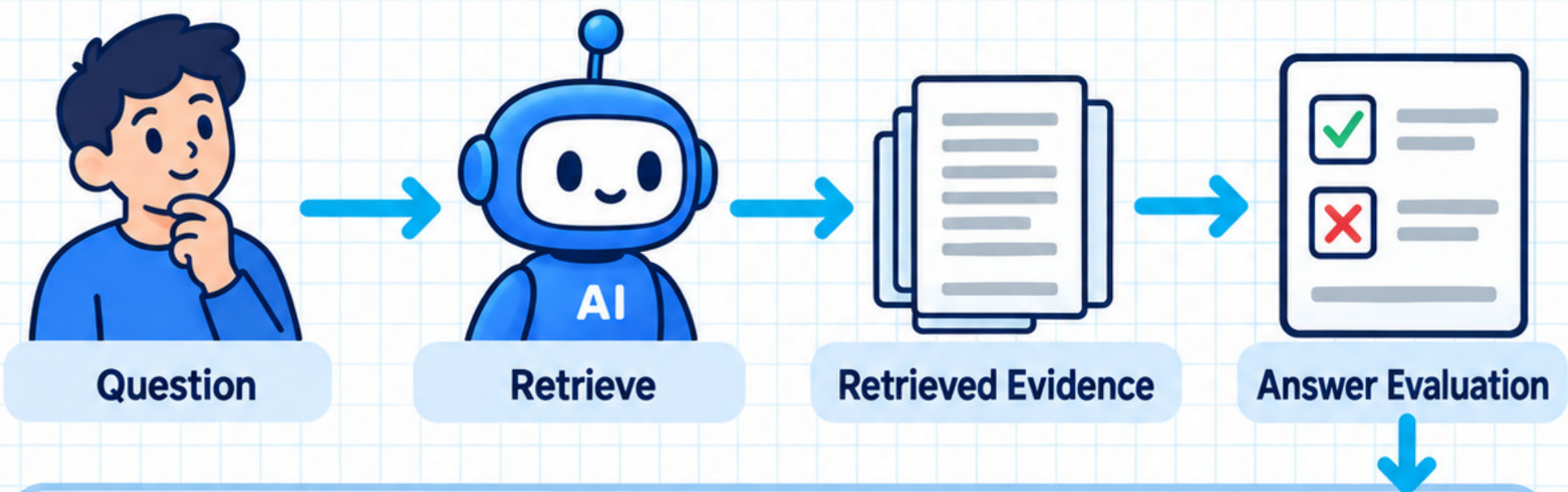
Compare fixed-size, structural and parent-child chunking.

Hold the other retrieval settings constant so the comparison is interpretable.



MEASURE MORE THAN HITS

Check whether retrieved evidence contains the required rule and exception.
Record answer correctness, returned tokens, chunk count and index size.
A keyword hit is not enough if the meaning is incomplete.



Evaluation Checklist

Item	Measurement
Rule and exception	
Answer correctness	
Returned tokens	
Chunk count	
Index size	

TOOLS HAVE SEPARATE JOBS



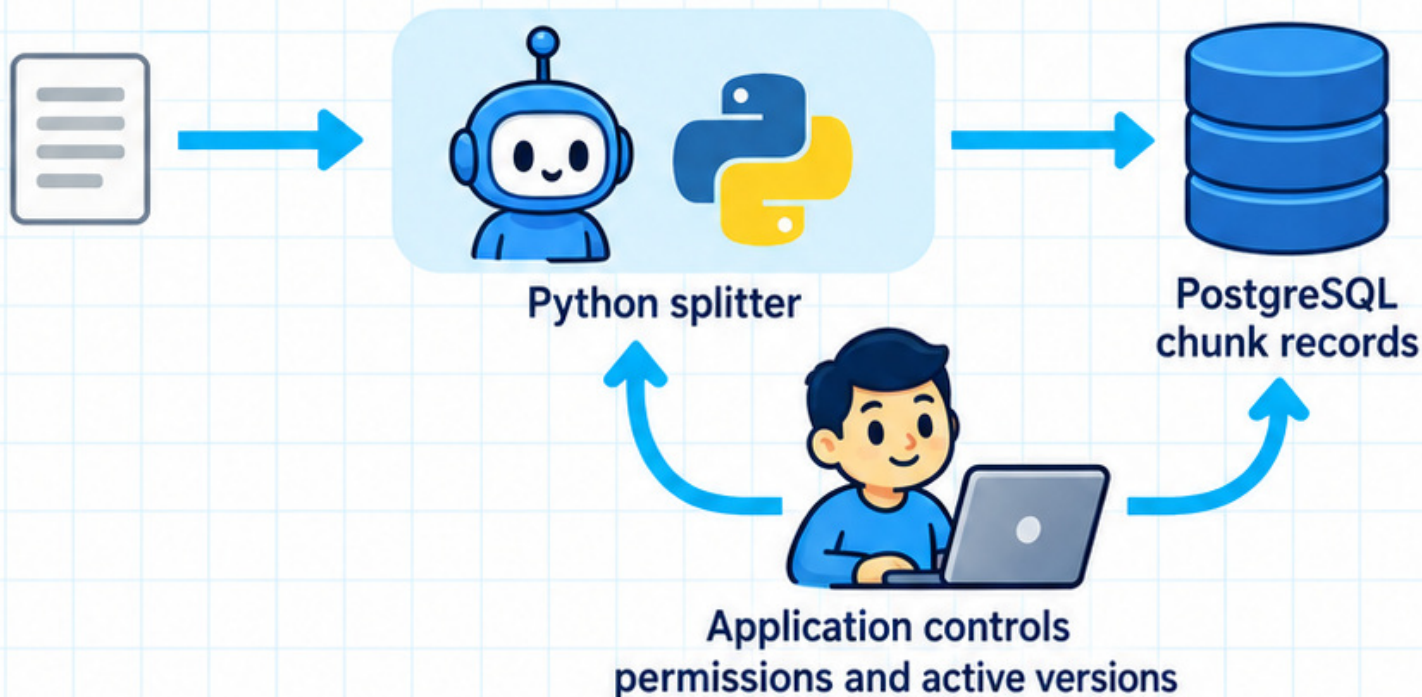
Python applies and versions the splitting rules.



PostgreSQL stores chunk records, metadata and source links.



A splitter library can help with boundaries; your application owns permissions and activation.



TRY THE BOUNDARY TEST

Toy source

Toy source: Unopened items: 30 days.
Opened headphones: excluded unless defective.



Question: Can I return defective opened headphones?

Choose a chunk boundary



List the metadata needed to verify the exception.



YOUR CHUNK CONTRACT

Keep the meaning, location, permissions and version together.

Compare chunk variants on the same questions.

Next: build an index that can be updated without serving stale evidence.

