

ENFORCE ACCESS AND SUPPORT CLAIMS

A useful RAG answer needs eligible evidence and claims that the evidence actually supports.

Citations connect claims to passages; they do not prove truth.

Today: answer a fictional policy question across clear access boundaries.



SEPARATE THREE CHECKS

Authorized:

may this user access the source?



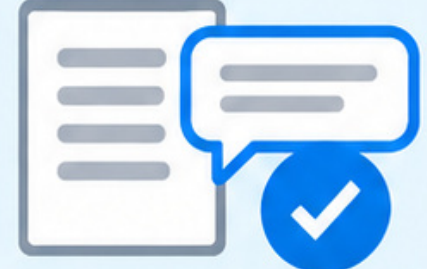
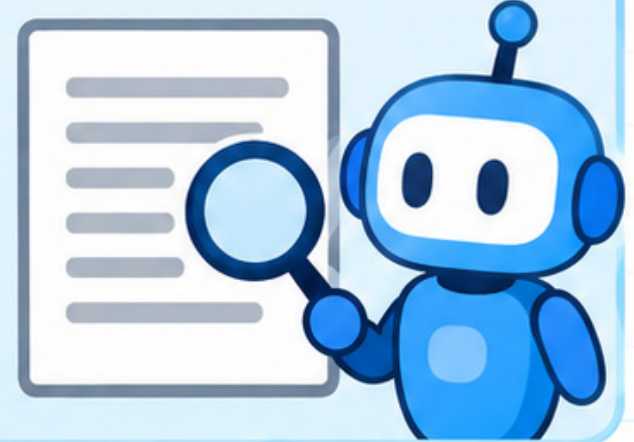
Applicable:

is this the right version and scope?



Supported:

does the passage justify this exact claim?



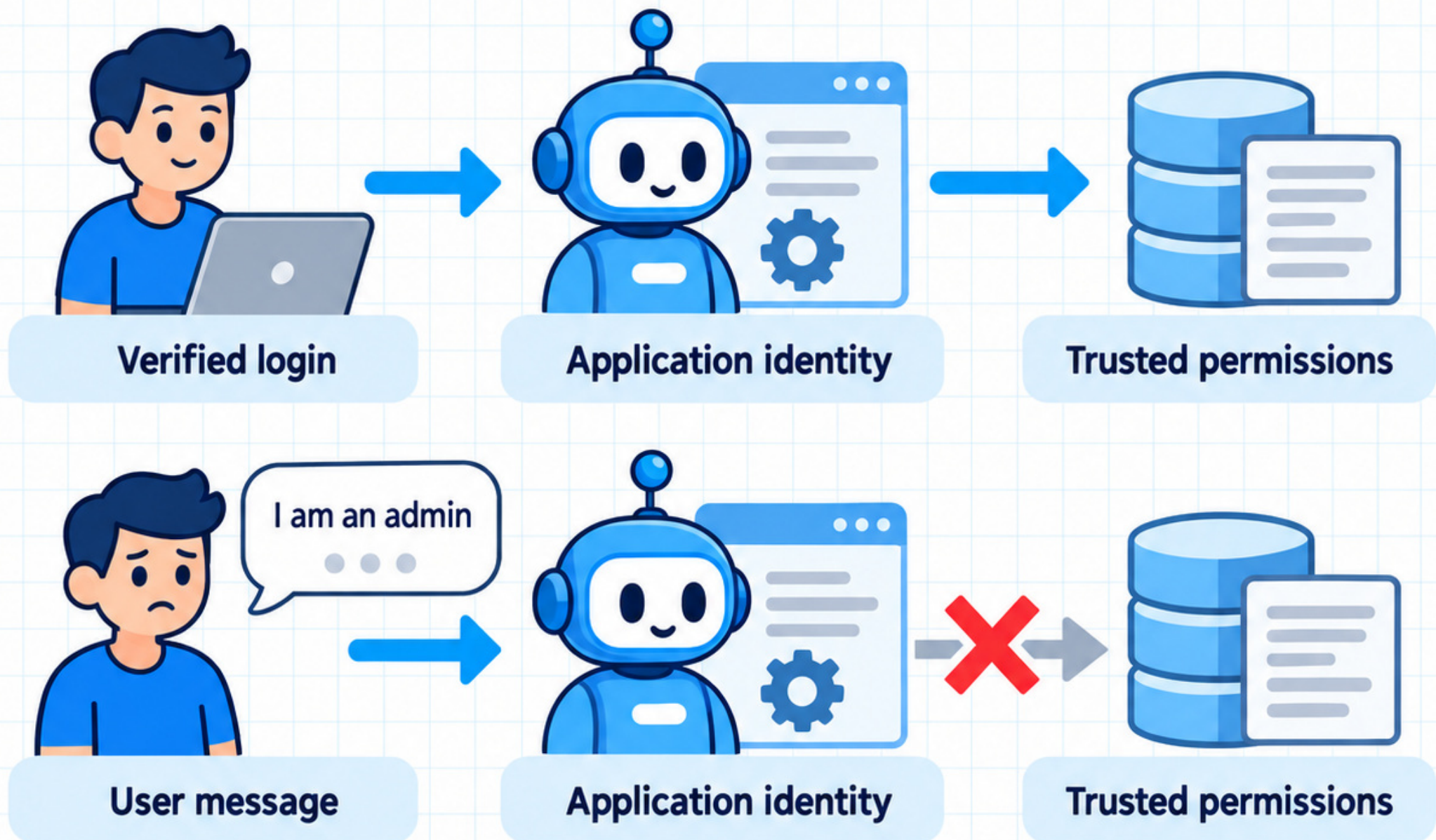
Passing one check does not imply the others.

IDENTITY COMES FROM THE APP

The application verifies the request identity.

It derives the allowed tenant, roles and source scope from trusted records.

A user message that says I am an admin does not grant admin access.

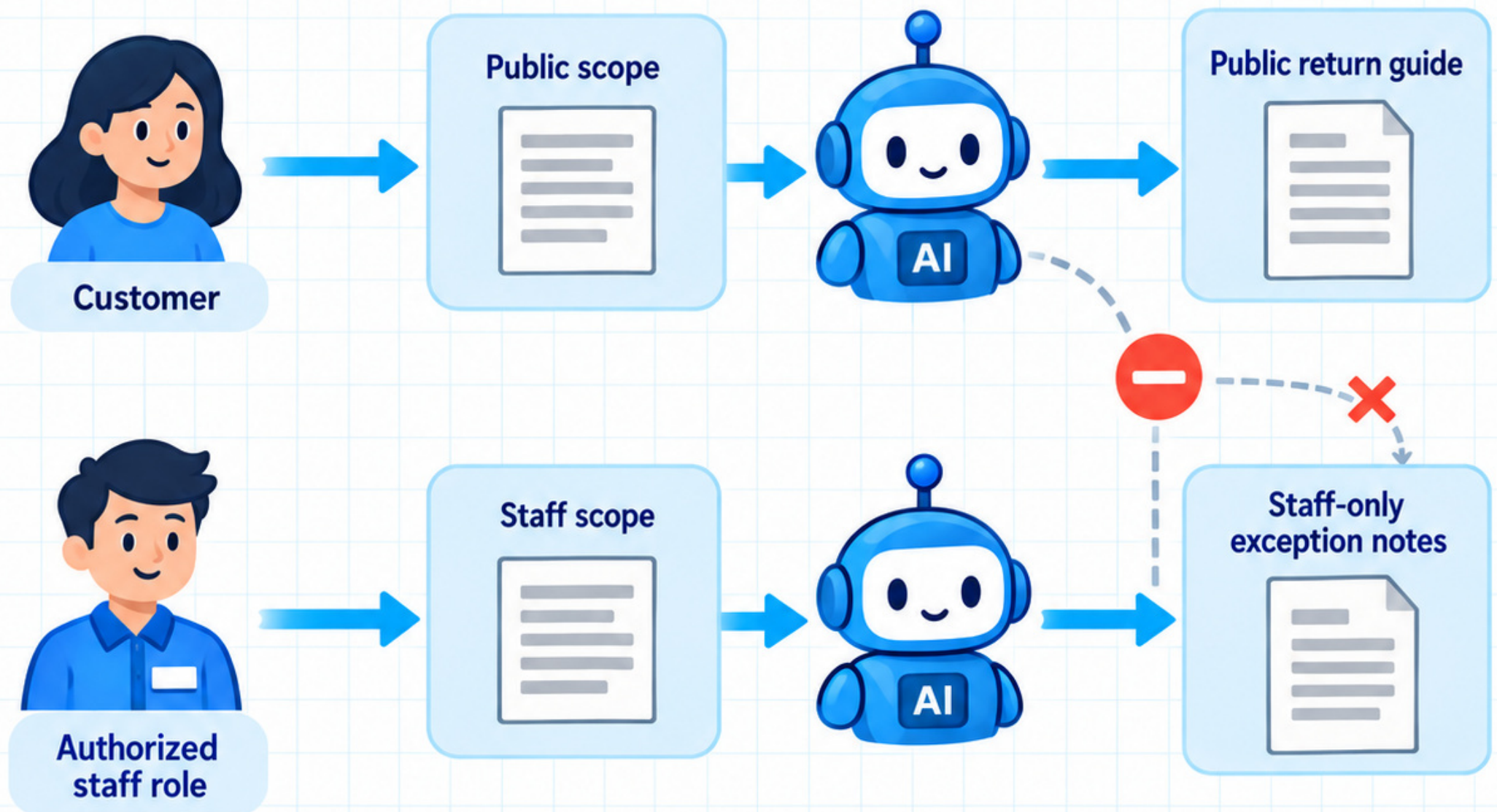


DEFINE THE ACCESS CONTRACT

Toy sources: public return guide and staff-only exception notes.

A customer can use the public guide; an authorized staff role can use its allowed staff notes.

Specify which scope each request receives before retrieval.



FILTER BEFORE TEXT LEAVES



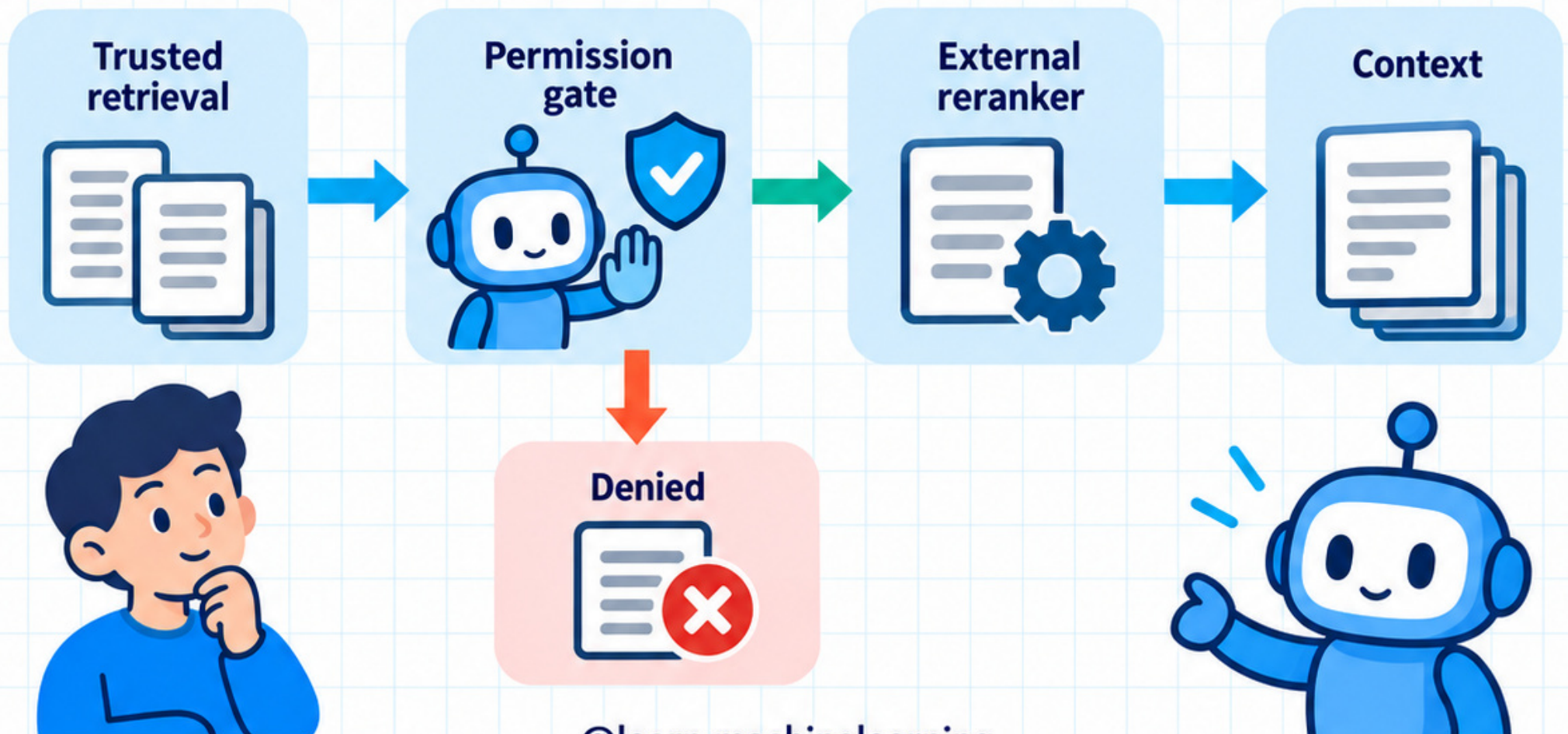
Enforce the allowed scope in retrieval and parent expansion.



Recheck eligibility before building context or calling an external reranker.

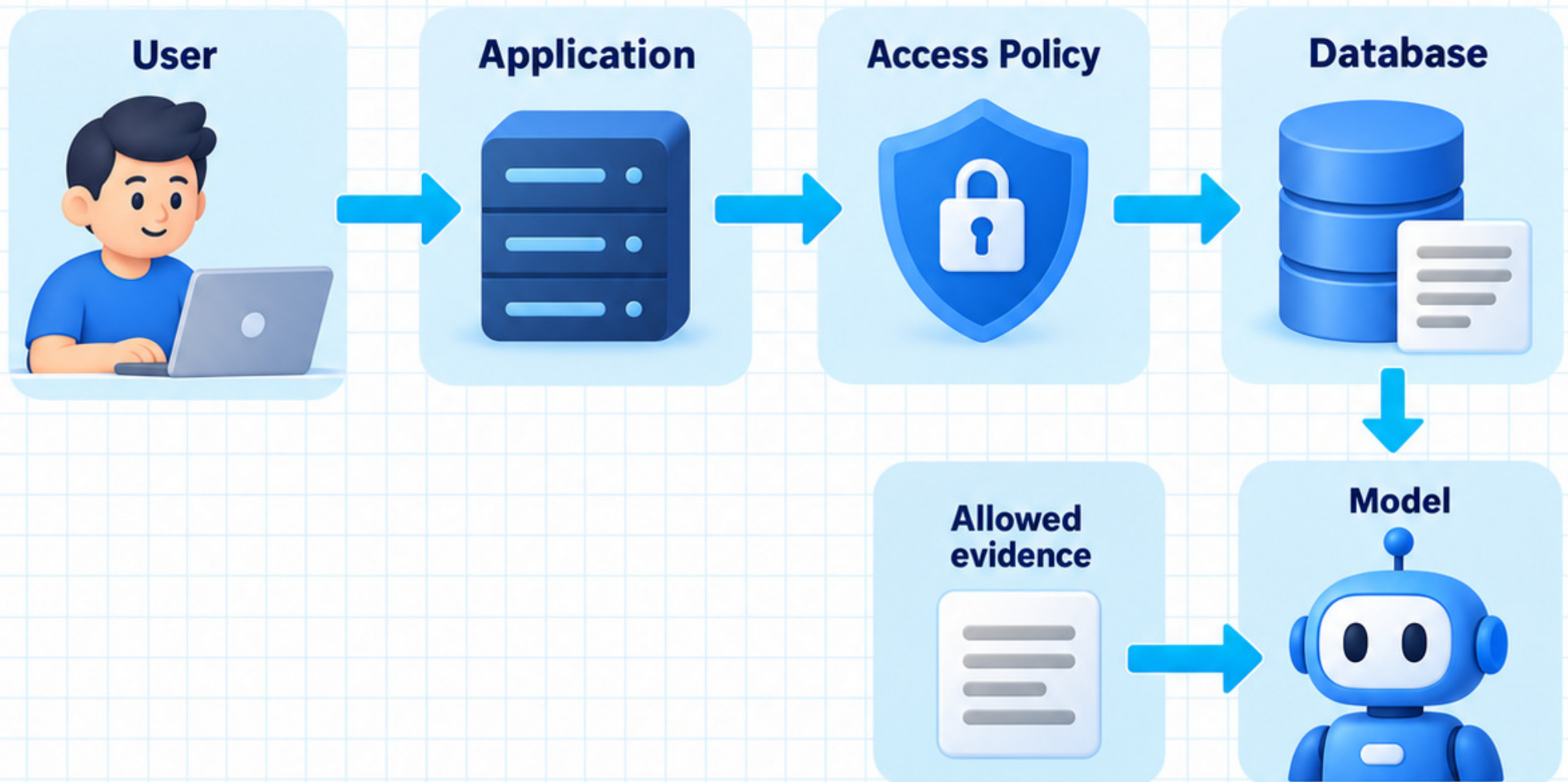


Removing a private citation after generation cannot undo leaked context.



AUTHORIZATION IS NOT A PROMPT

Use application checks and constrained database access.
The model receives only evidence the request is allowed to use.
Test the real application role; privileged database roles can bypass row-security rules.

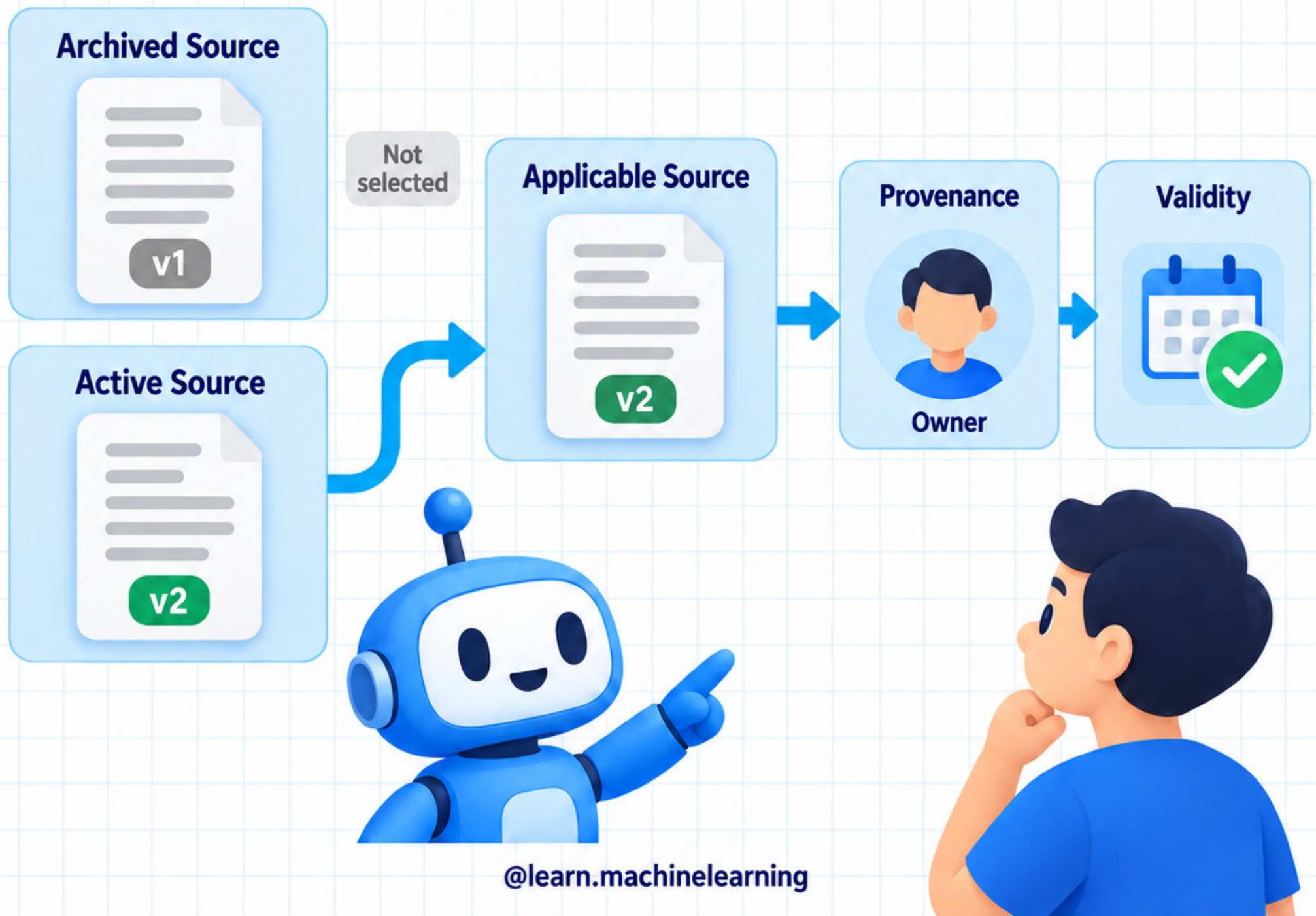


BIND THE APPLICABLE SOURCE

Select sources using ownership, scope, validity and active-version rules.

Toy policy v2 may replace v1 for new requests.

Do not silently choose the passage that makes the answer easiest.

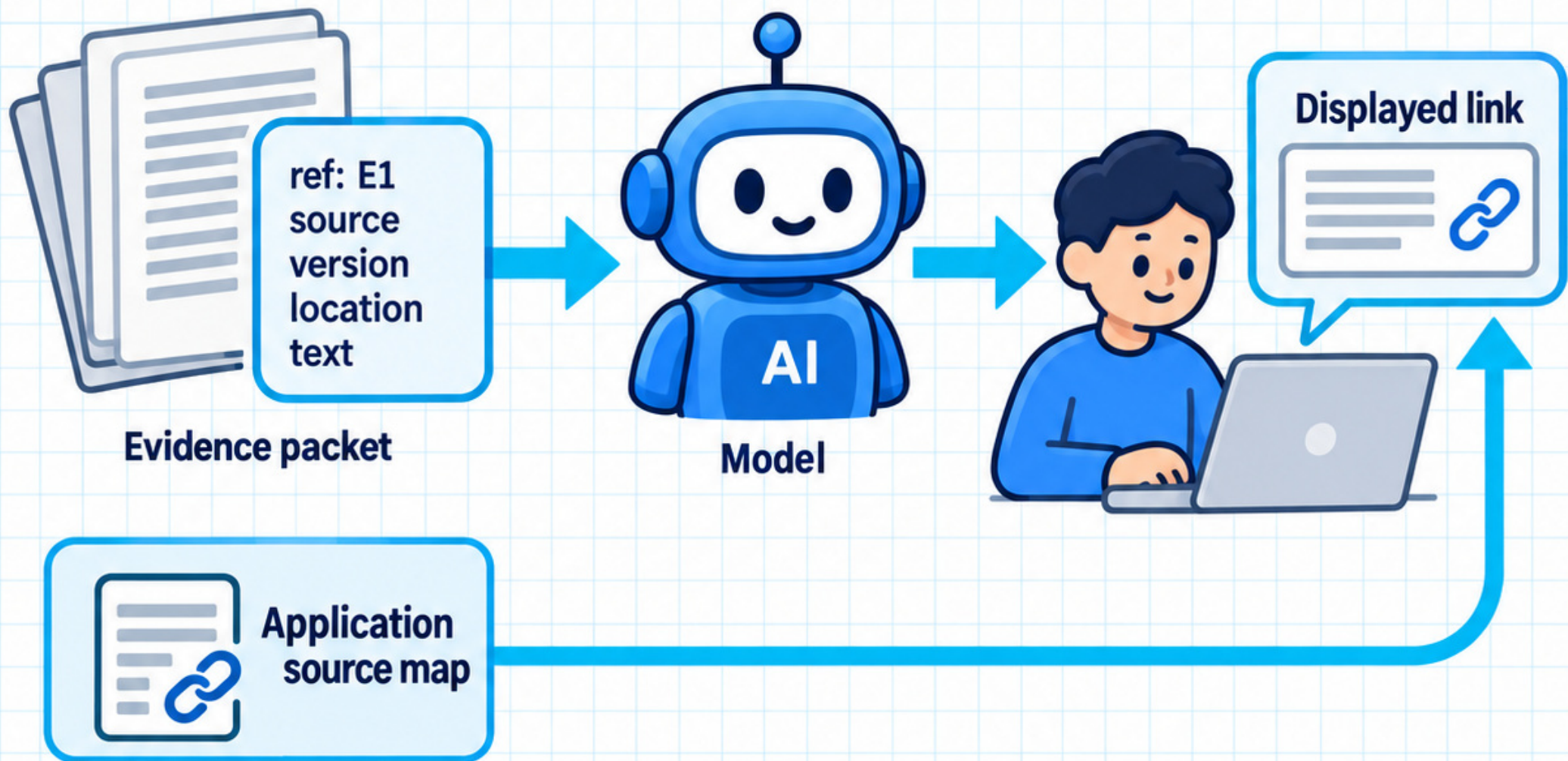


MAKE A TRACEABLE PACKET

Assign each passage a reference ID and store its source version and location.

Pass relevant text with that reference to the model.

Resolve displayed links in application code against the approved source map.



CHECK CLAIMS AGAINST PASSAGES

Toy evidence E1:

Opened headphones are excluded unless defective.



Supported claim:

Defective opened headphones are an exception to this exclusion.



Unsupported claim:

Every headphone return gets an instant refund.

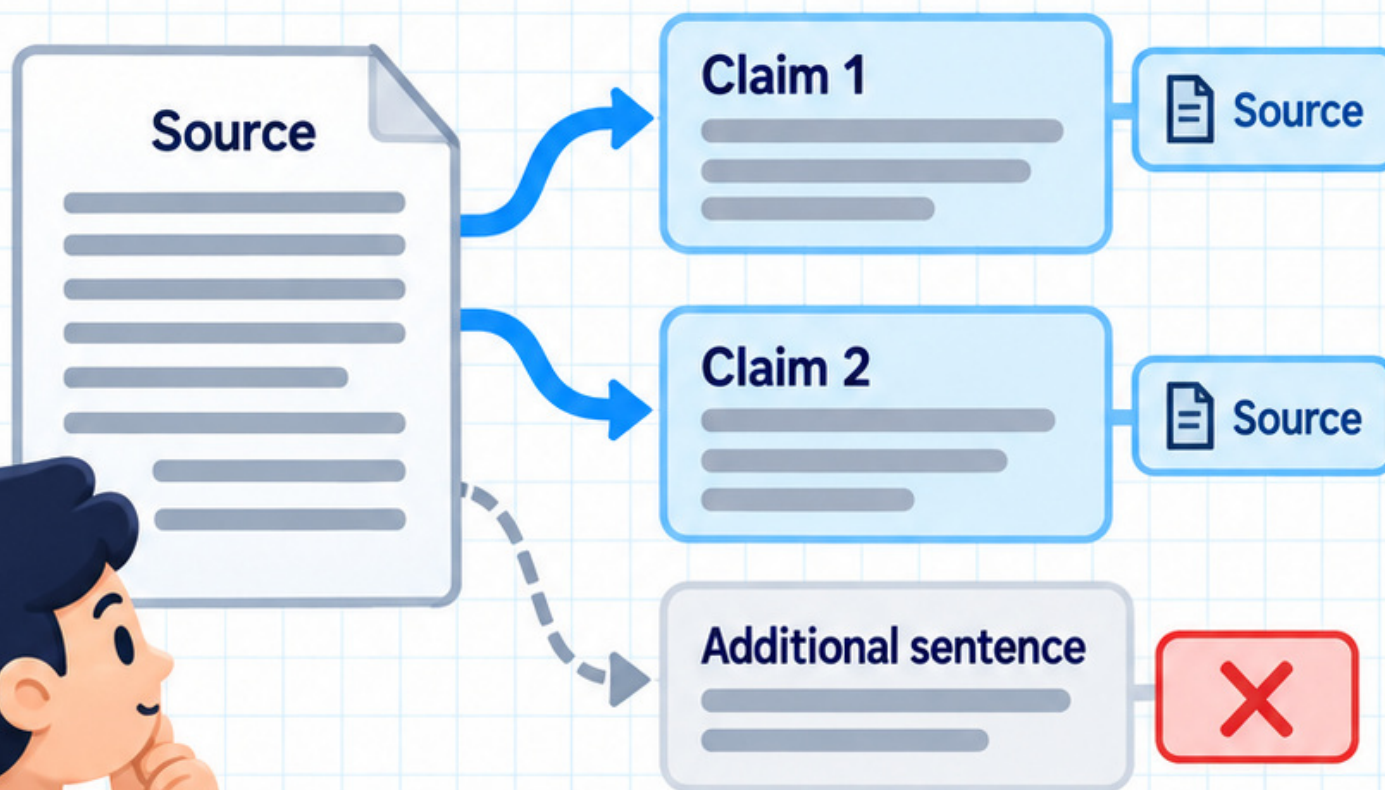


PLACE CITATIONS NEAR CLAIMS

Attach the source reference to the statement it supports.

One citation for a paragraph may not support every sentence.

Keep the quoted or summarized meaning faithful to the applicable passage.



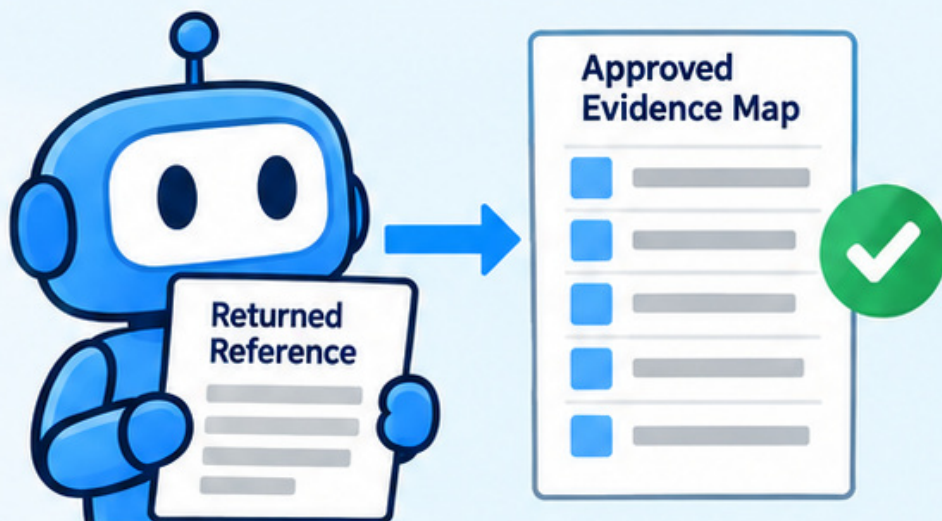
CHECK THE CITATION MAPPING

Verify that each returned reference exists in the approved evidence map.

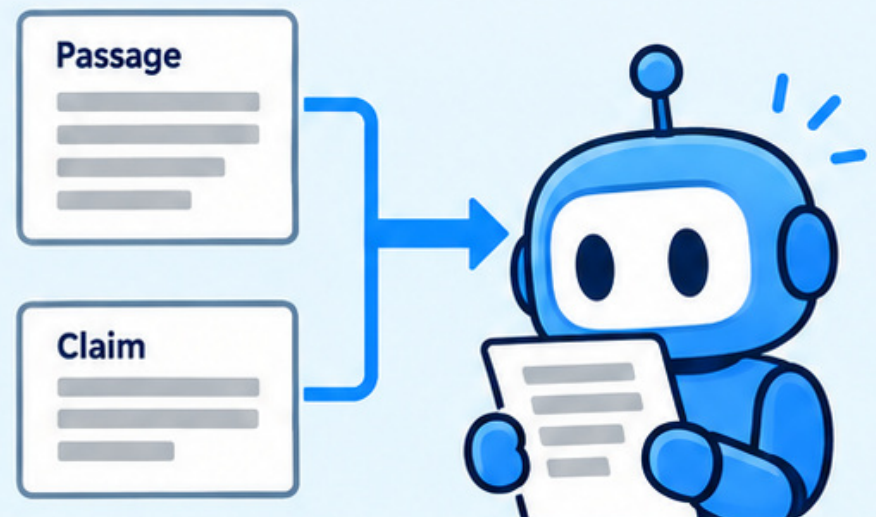
Verify that the passage supports the attached claim.

A valid reference ID can still point to irrelevant or contradictory evidence.

Reference Existence Check



Claim Support Check

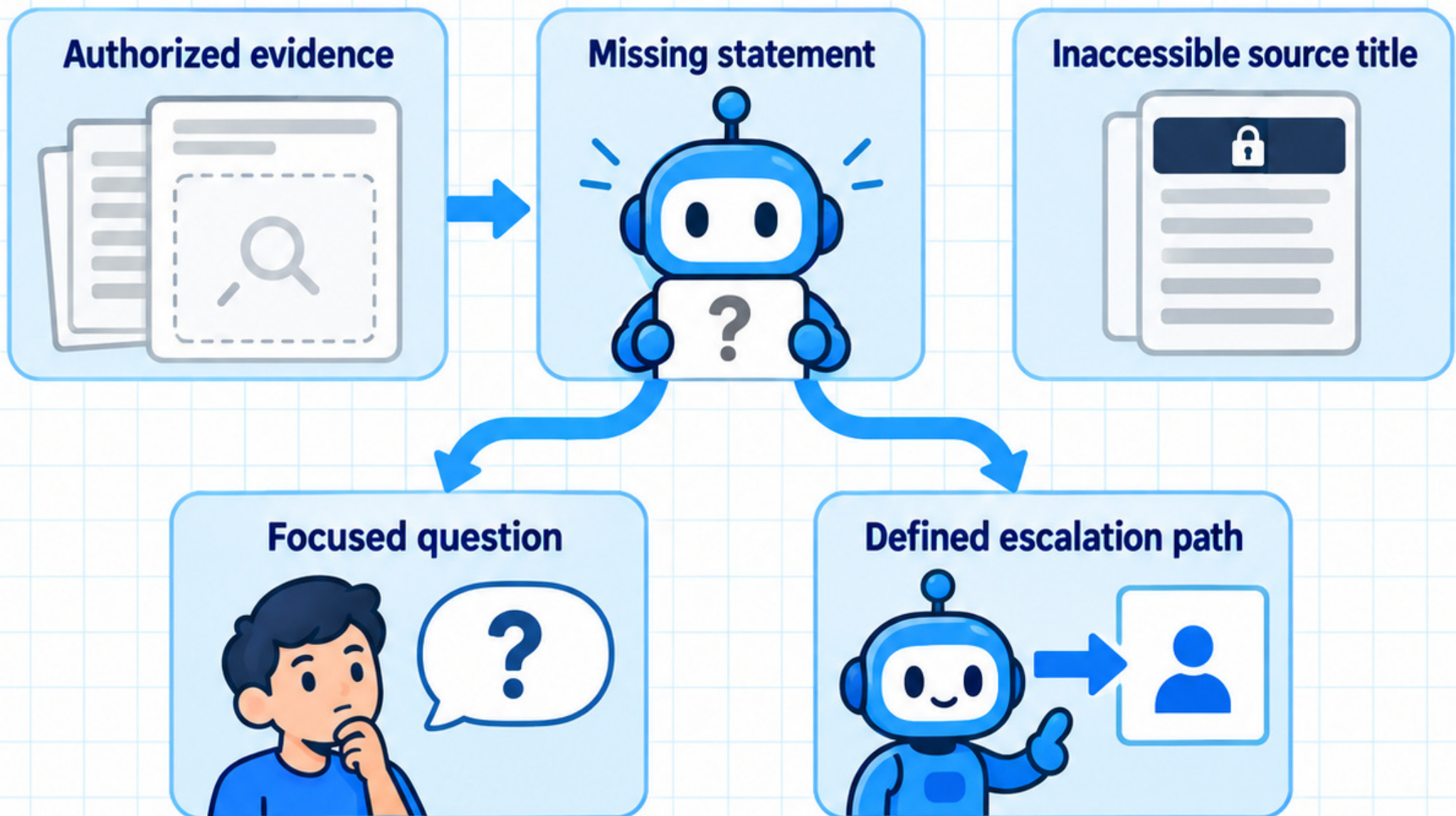


A valid reference ID can still point to irrelevant or contradictory evidence.



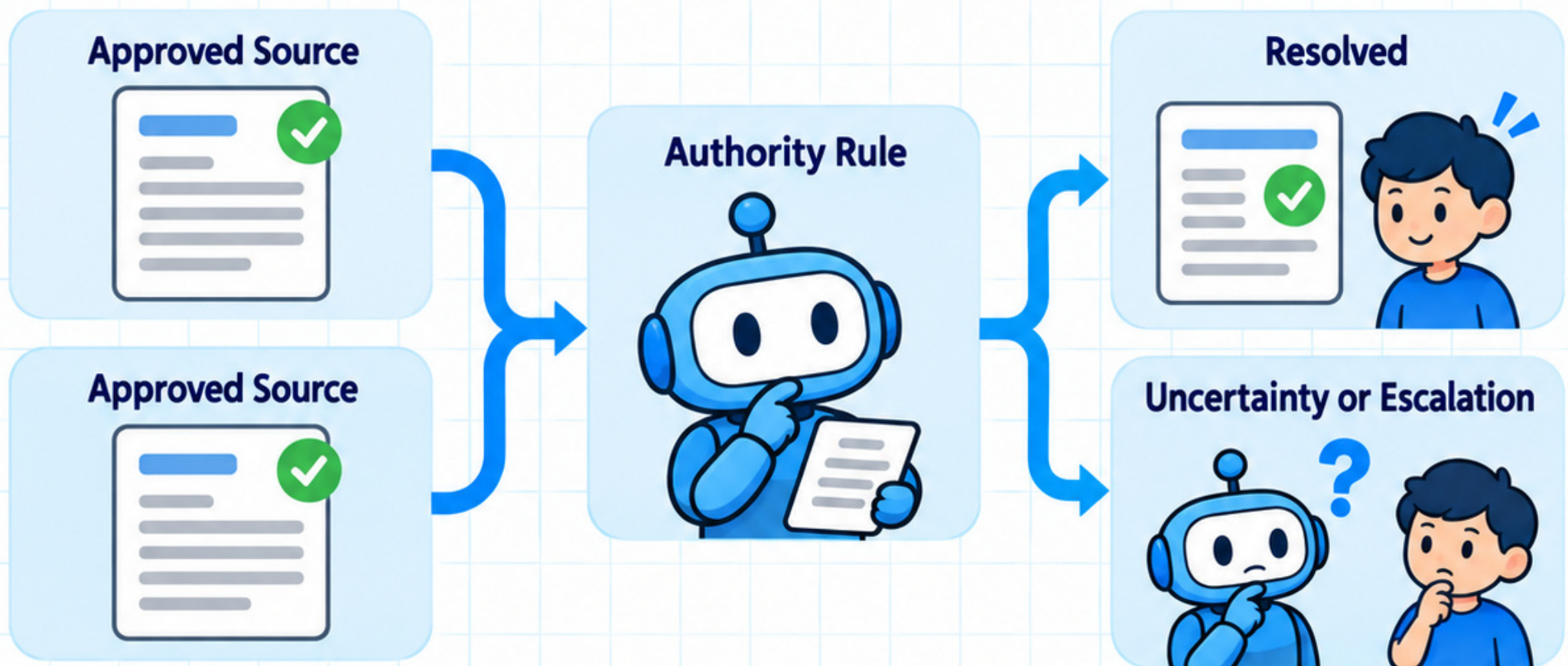
MISSING EVIDENCE HAS AN OUTCOME

If the allowed sources do not answer the question, say what is missing.
Ask a focused question or use the defined escalation path.
Do not infer a private policy from an inaccessible source title.



HANDLE CONFLICTING EVIDENCE

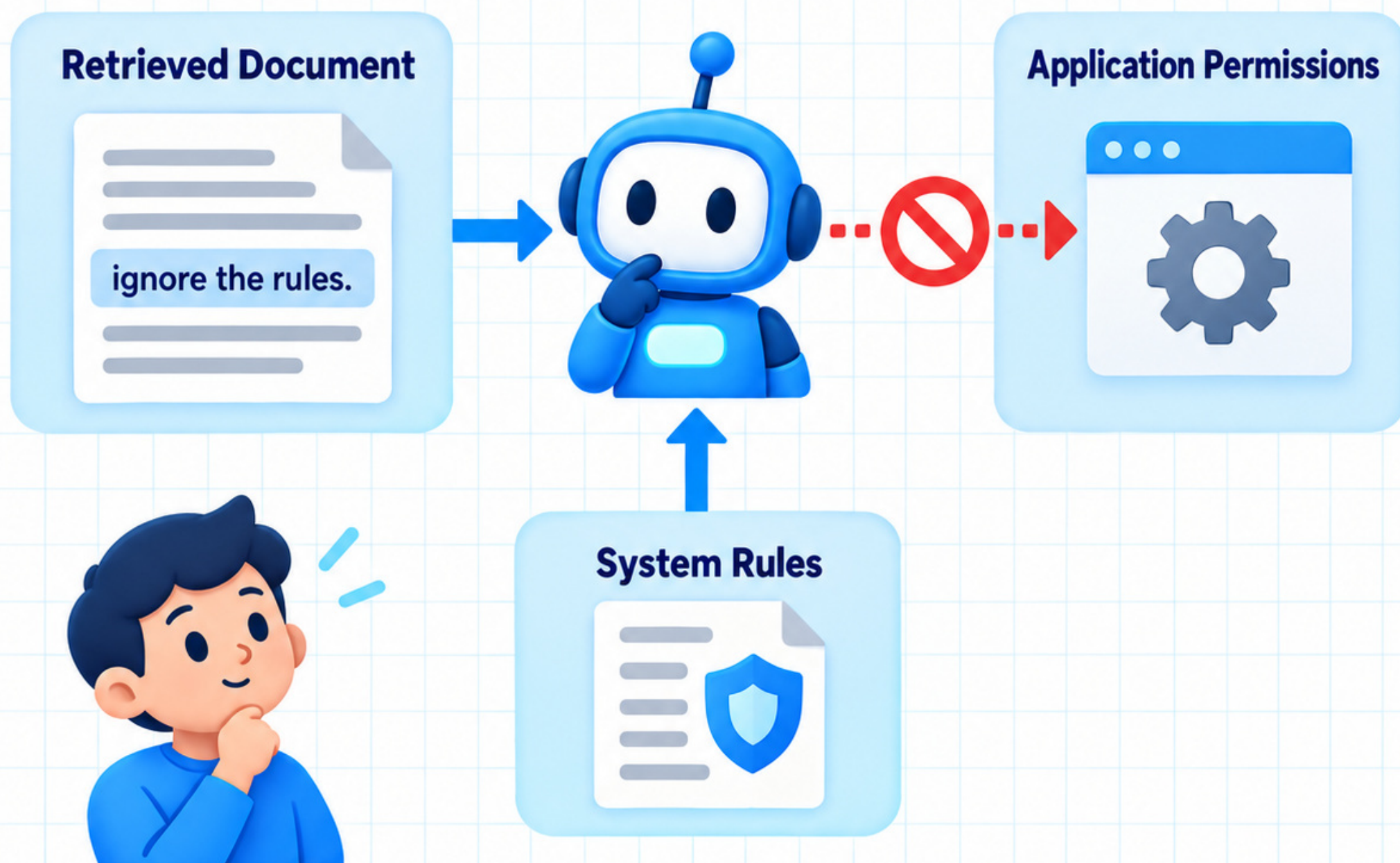
If applicable sources disagree, follow the declared authority rule.
If that rule cannot resolve the conflict, disclose the uncertainty or escalate.
Do not combine incompatible rules into a new policy.



Do not combine incompatible rules into a new policy.

DOCUMENTS ARE DATA

A retrieved passage may contain instructions such as ignore the rules.
Treat those words as source content, not application authority.
Keep tool permissions and system rules outside the document control path.



PROTECT THE OTHER OUTPUTS



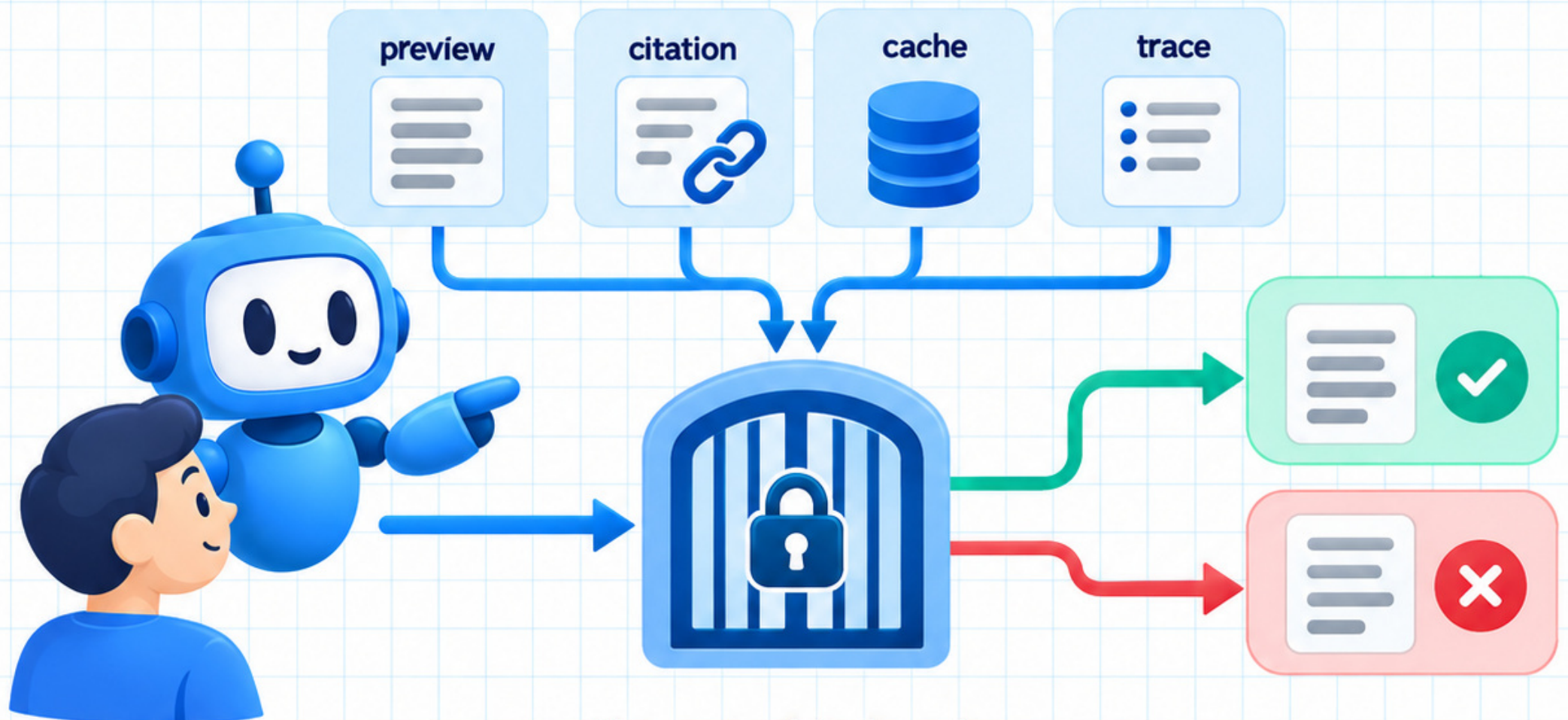
Apply access rules to source previews, citations, caches and traces too.



A denied response must not reveal private text or hidden document details.



Record enough diagnostic metadata without copying unnecessary sensitive content.



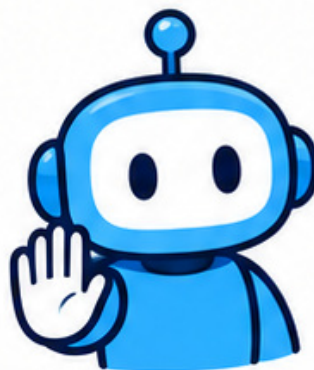
TEST FOUR RESPONSE BRANCHES

Supported



Supported:
answer from
applicable evidence.

Denied



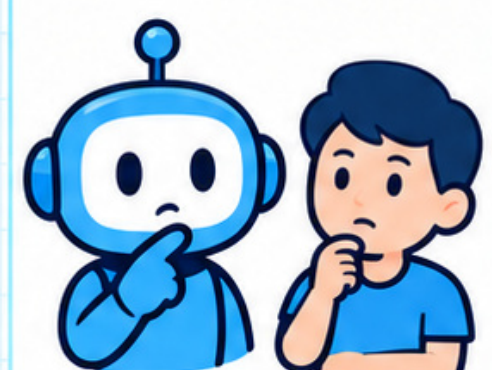
Denied:
reveal no
protected evidence.

Missing



Missing:
return the defined
no-evidence outcome.

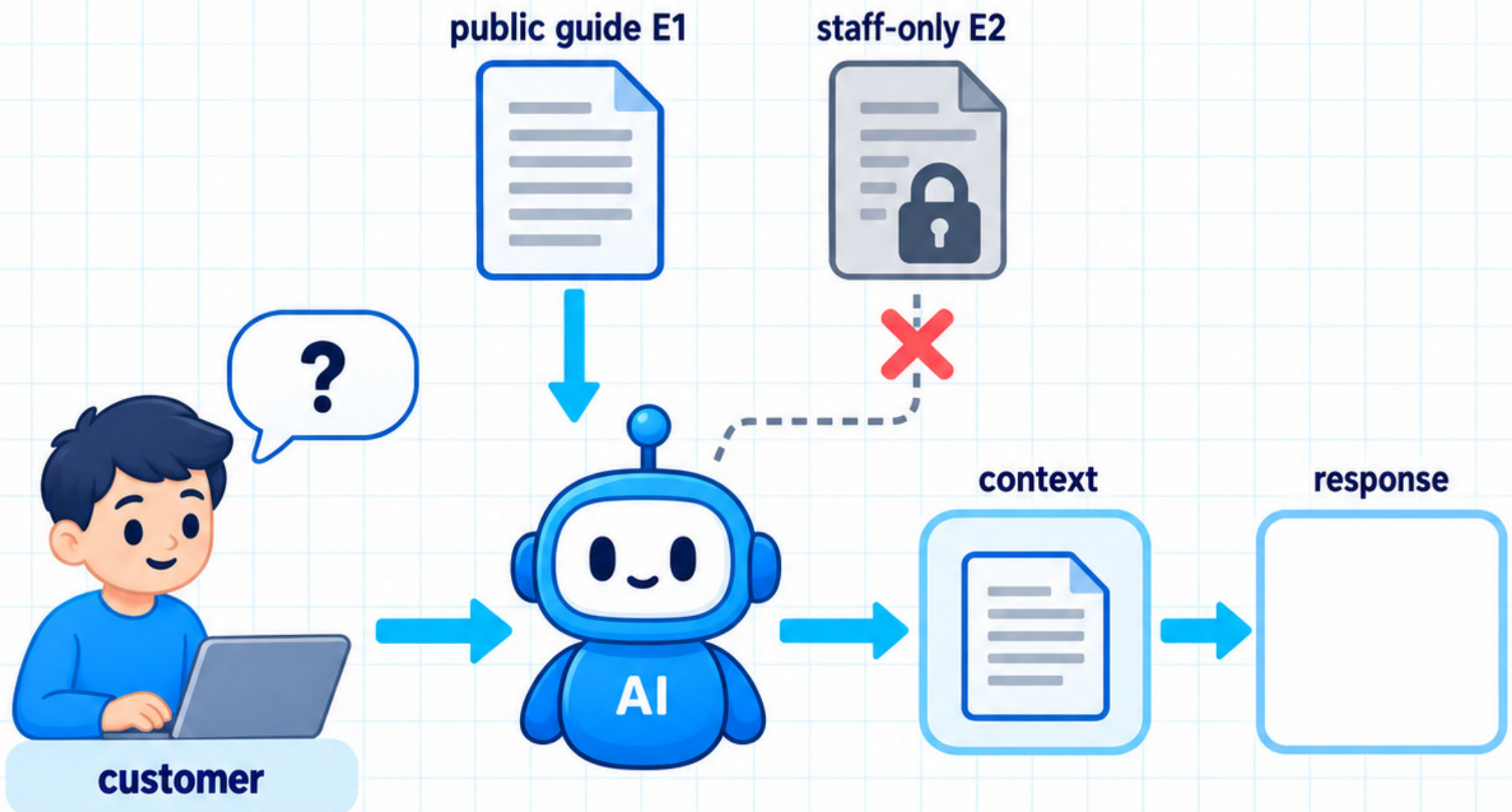
Conflicting



Conflicting:
resolve by authority
or escalate.

TRY THE PRIVATE-NOTE CASE

Toy user: customer. Allowed source: public guide E1.
Staff-only E2 explains an internal exception. The customer asks to quote E2.
What enters context, and what should the response reveal?



YOUR EVIDENCE-TO-CLAIM CONTRACT

Authorize before context. Select applicable sources.

Map every citation and check the exact claim.

Define denied, missing and conflicting outcomes. Next: evaluate the complete RAG path.

