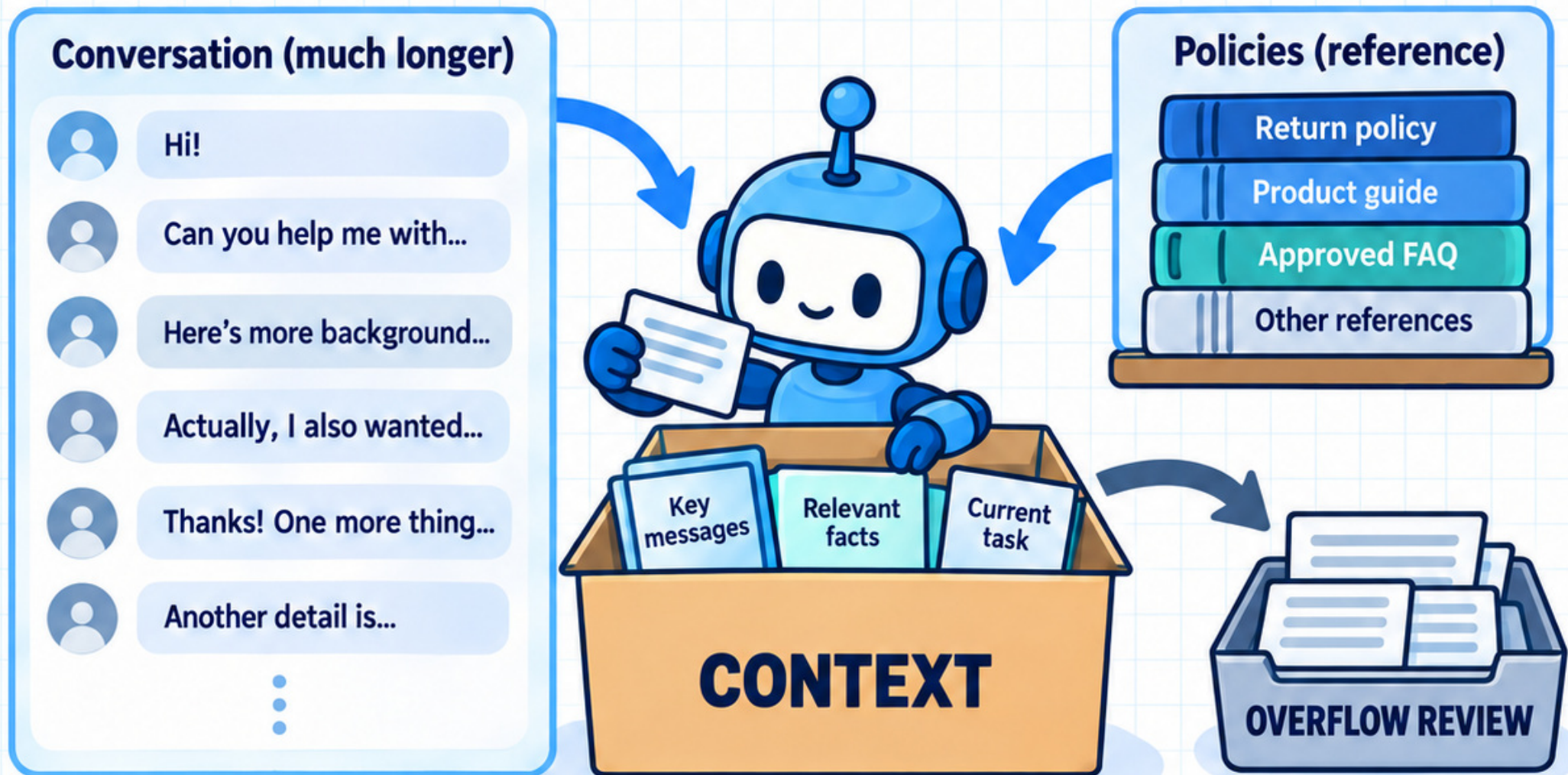


MORE CONTEXT IS NOT A PLAN

A long conversation can bury the fact that changes the answer. Choose what enters the model, what stays outside, and what happens on overflow.



Today: design a context budget.

NAME THE CONTEXT PIECES

INSTRUCTIONS

Instructions define the task.



Example:
Answer in a clear,
concise style.

LATEST REQUEST

The latest request defines the need.



Example:
Can I return this item?

AUTHORIZED EVIDENCE

Authorized evidence supports the answer.



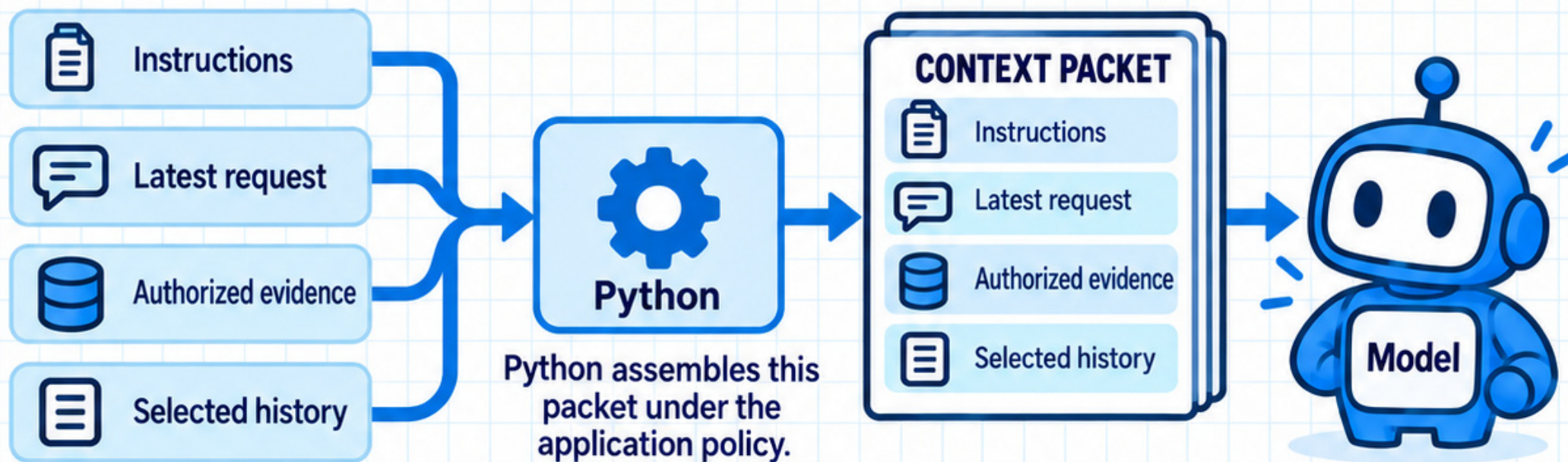
Example:
Approved policy:
final-sale items are
excluded.

SELECTED HISTORY

Selected history explains prior decisions.



Example:
Earlier decision:
user wants a return.

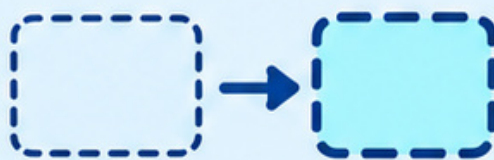


Each piece needs a reason to be included.

THE WINDOW IS A LIMIT



A model has an input/output capacity contract.

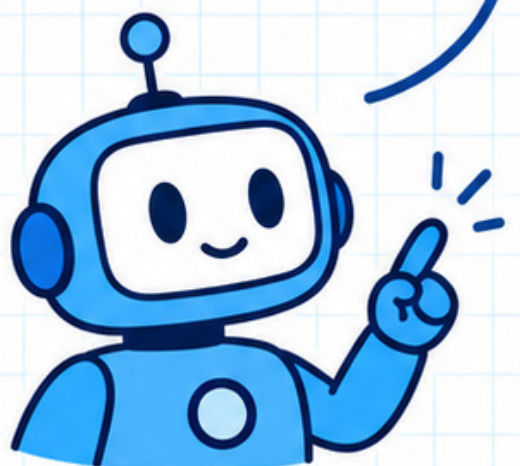
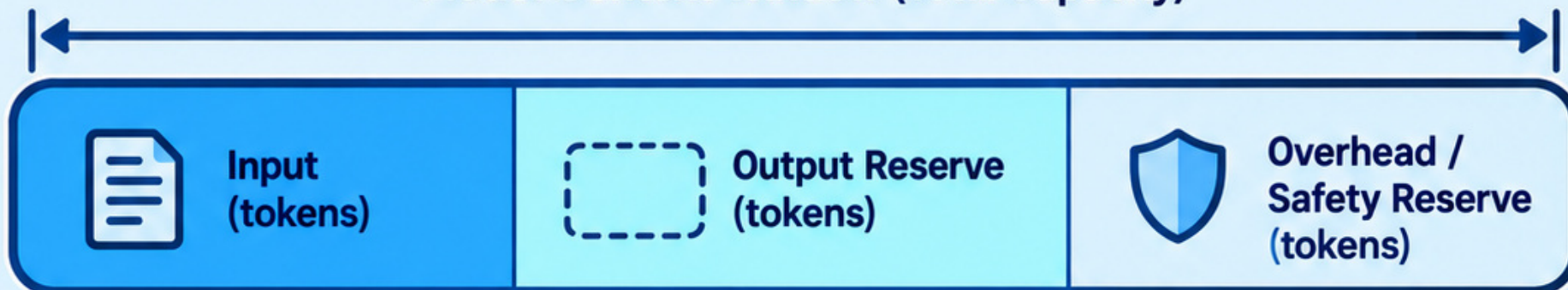


Reserve space for the expected output and provider-specific overhead.



Verify limits and token accounting for the chosen model and modality.

Model Context Window (Total Capacity)



The window is a fixed limit. Plan your input, reserve output space, and include overhead so you stay within capacity.

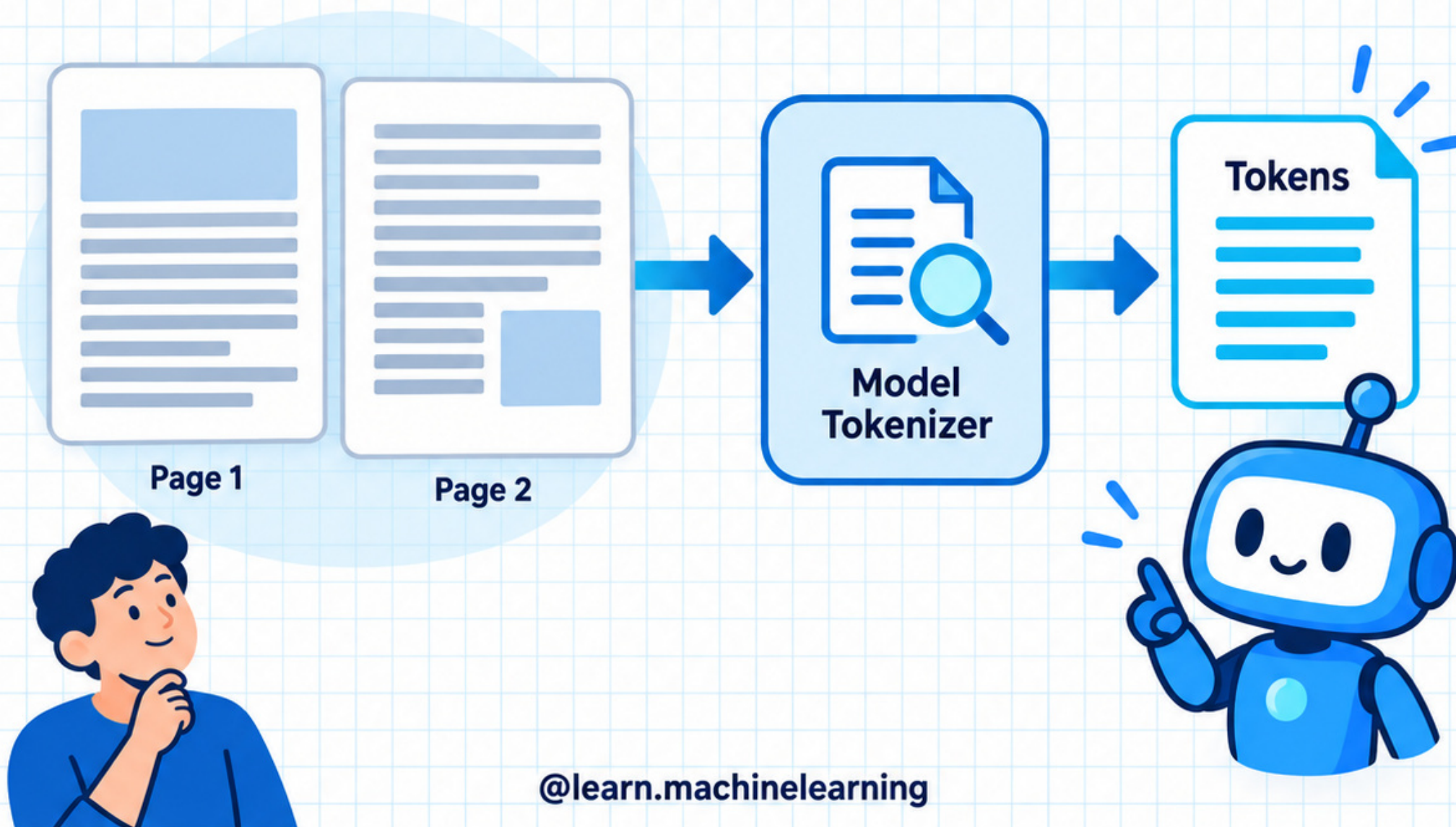


COUNT TOKENS, NOT PAGES

Tokens are model-specific units.

A page count or character count is only a rough proxy.

Use the selected model tokenizer or provider accounting, and log actual usage when available.



A TOY BUDGET

**Illustrative total allowance:
8,000 tokens.**

**Reserve 1,000 for output and
500 for overhead and safety.**

**Input packet target:
at most 6,500 tokens.**

**This is a teaching allocation,
not a model specification.**

**Total (toy)
8,000 tokens**

500 tokens

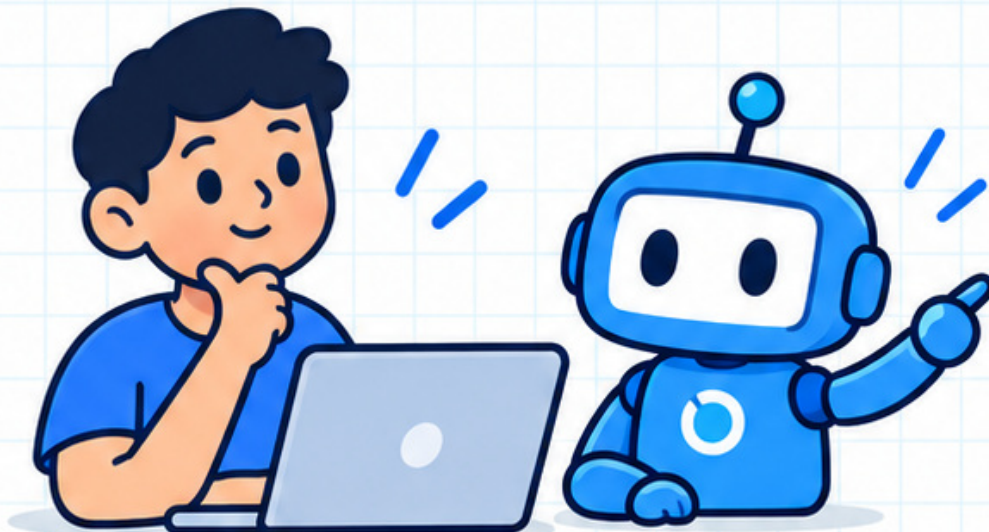
← **Overhead
and safety**

1,000 tokens

← **Output**

6,500 tokens

← **Input packet
(at most)**



ALLOCATE BY THE TASK

**Toy input allocation: instructions 500; request 500; evidence 3,500; selected history 2,000.
Total: 6,500 tokens.**

Instructions



500 tokens



Unused
capacity
(up to 500)

Request



500 tokens



Unused
capacity
(up to 500)

Evidence



3,500 tokens



Unused
capacity
(up to 3,500)

**Selected
history**



2,000 tokens



Unused
capacity
(up to 2,000)

Total: 6,500 tokens.

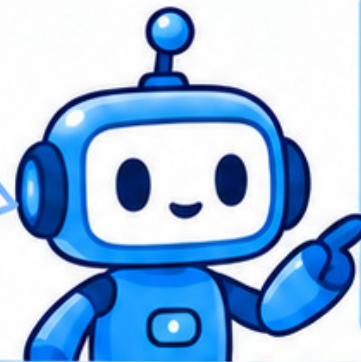
**These are maximum allocations,
not a requirement to fill every slot.**

PROTECT REQUIRED FACTS

- 1 For a return answer, keep the deadline, applicable region, exception, and policy version.
- 2 Rank evidence by relevance and authority.
- 3 Remove duplicated text before removing a condition that changes the answer.

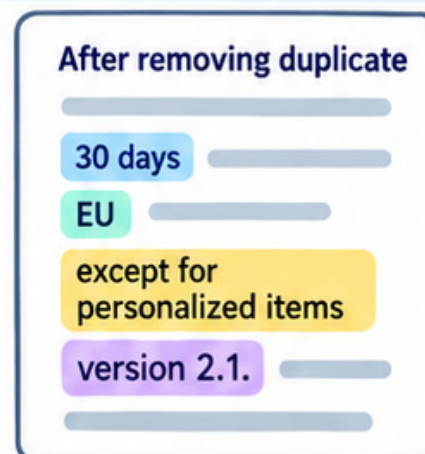
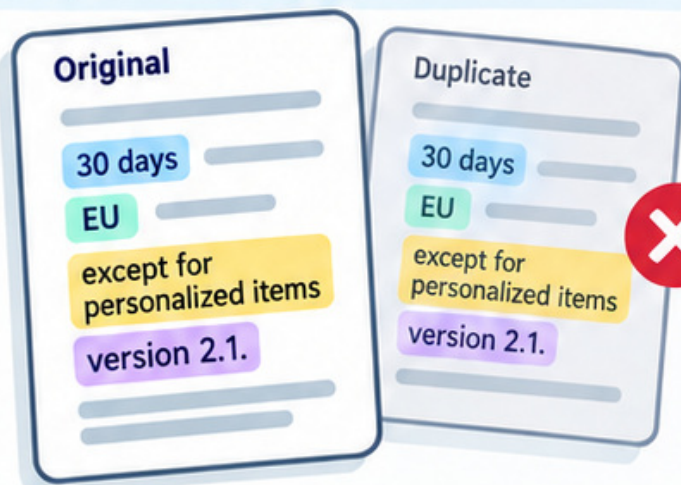
Toy policy example: preserve all four facts.

Returns must be requested within 30 days in the EU, except for personalized items, according to policy version 2.1.



30 days	= deadline
EU	= region
except for personalized items	= exception
version 2.1.	= policy version

Duplicate copy is removed while original facts stay.



Original facts stay.



HISTORY IS NOT CURRENT TRUTH

Toy conversation and records

An earlier message may contain an old address, obsolete policy, or mistaken claim.

Use current authorized records for current facts.

Keep unresolved questions and decisions that still affect the task.

Conversation History



My address is 123 Maple St.



I think the policy allows this.



Can you help me with this?

Relevant prior intent

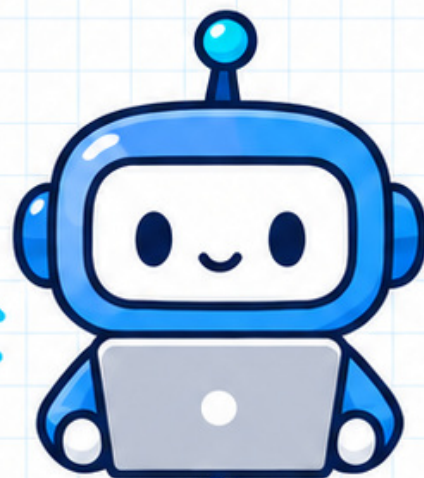
Current Authorized Records



Address:
456 Oak Ave.

Policy:
Updated version

Current facts from records

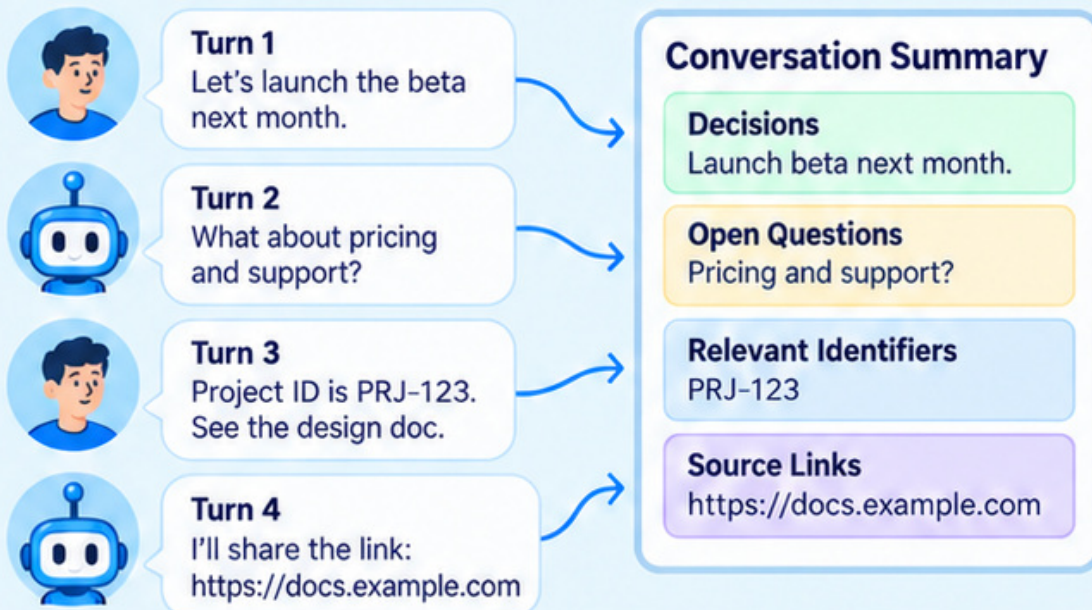


Answer using current records

SUMMARIES NEED A CONTRACT

Toy conversation summary

A conversation summary should preserve decisions, open questions, relevant identifiers, and source links.



Record its version and covered turns.

Summary (v1)

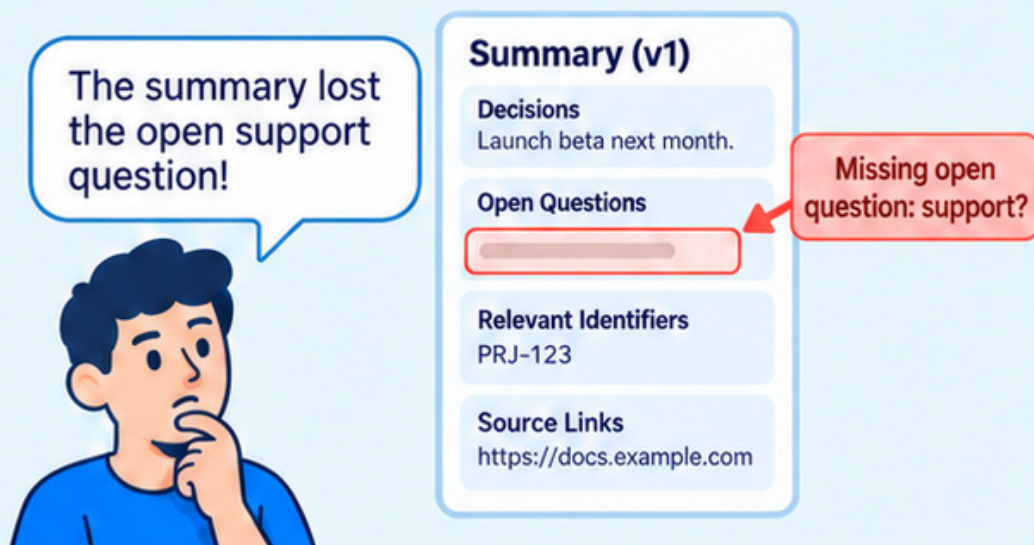
Version: v1

Covered turns: 1-4

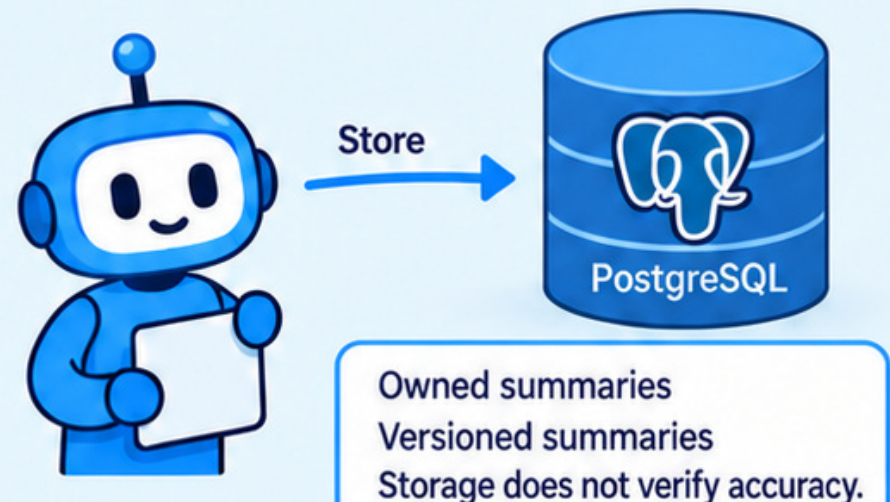
Content:



A generated summary can omit or distort facts; check important ones.



PostgreSQL can store owned, versioned summaries; storage does not verify accuracy.

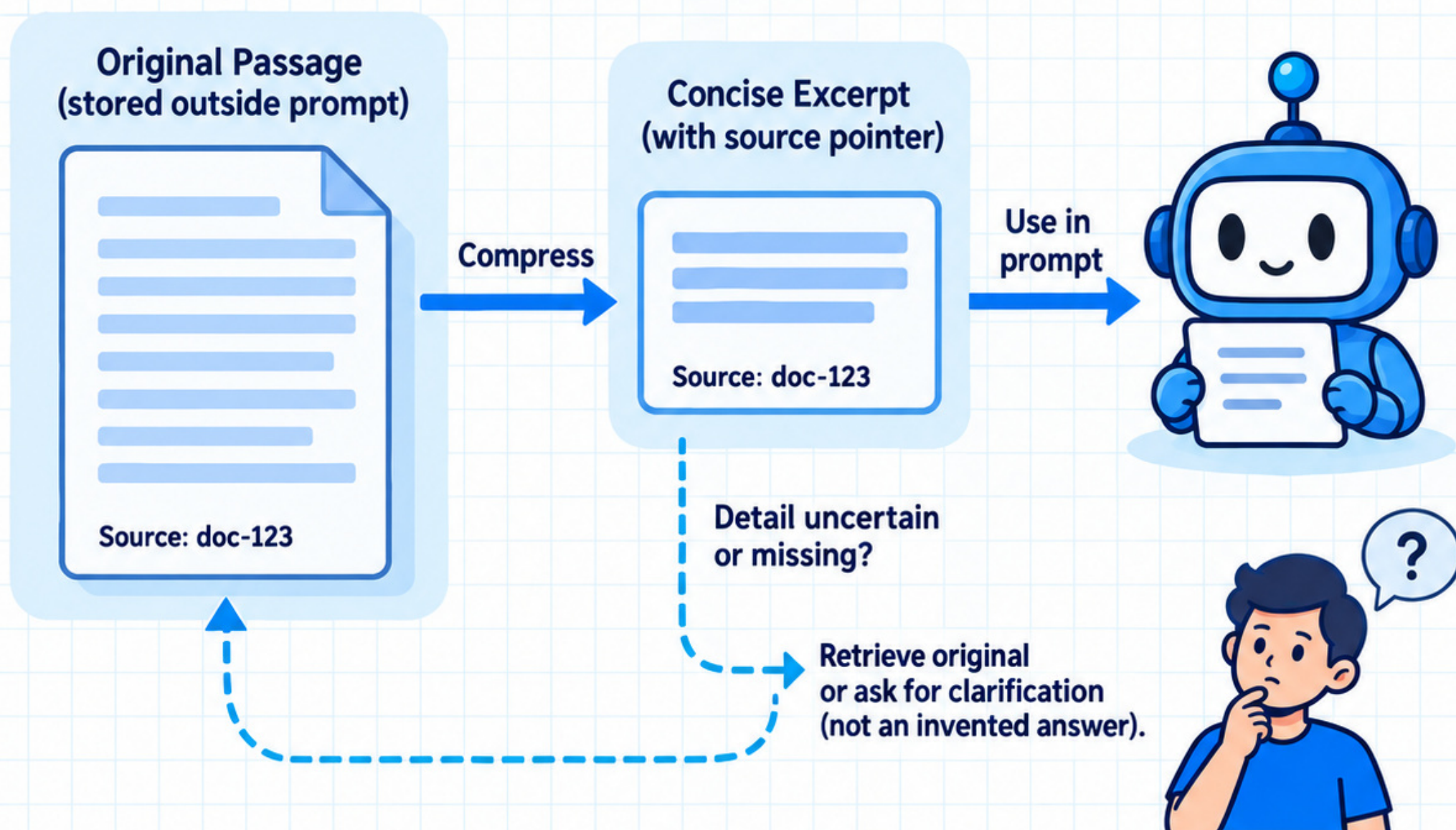


COMPRESS WITH TRACEABILITY

Keep the original evidence outside the prompt.

A compressed passage should point back to its source and preserve required conditions.

If an important detail is uncertain, retrieve the original or ask for clarification.



IMAGES NEED THEIR OWN POLICY



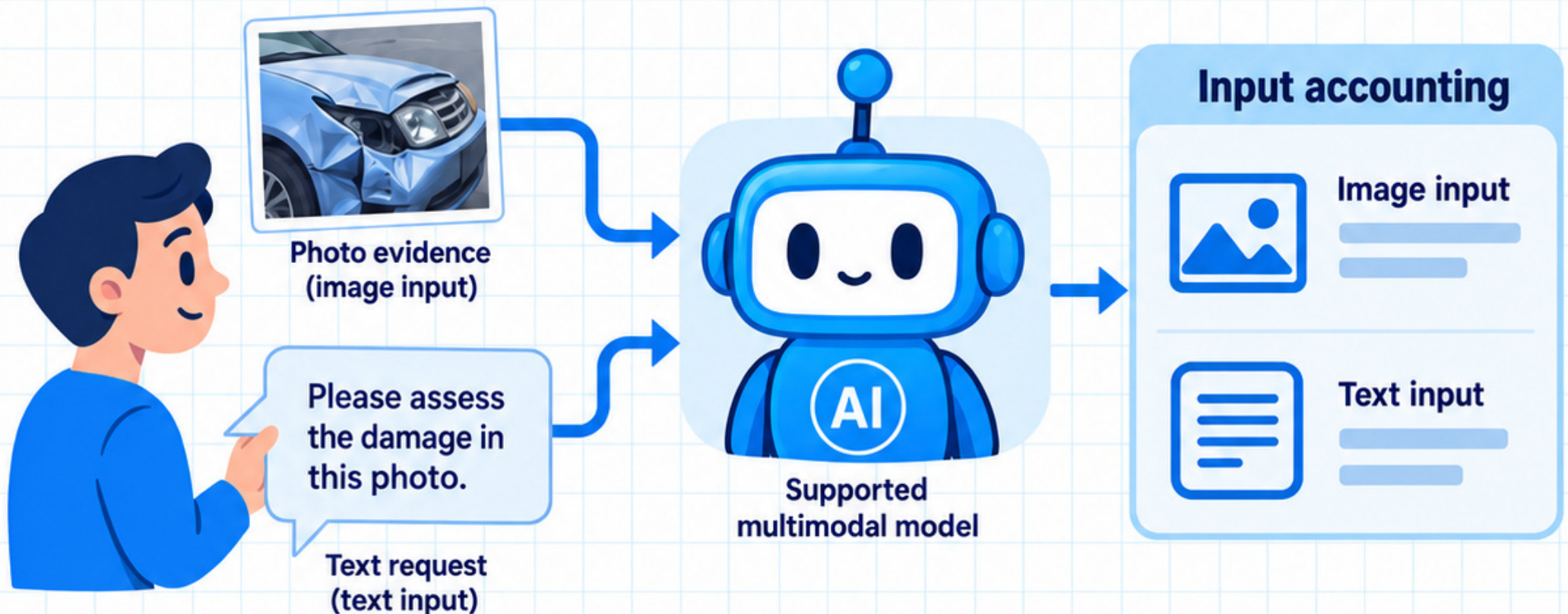
Toy task: inspect a damage photo.



Use an authorized image and a model that supports the required modality.



Account for image processing limits and cost; text token counts alone may be insufficient.



A PHOTO HAS LIMITS

Toy damage and order example

A damage photo may show a broken item.
It does not prove purchase ownership, order status, or refund eligibility.
Resolve those facts through authorized records and the policy.

Photo

Shows visible damage.



Authorized Order Record

Confirms ownership.



Order #12345

Item: Mug

Date: Jan 10, 2024

Status: Delivered

Customer: Alex

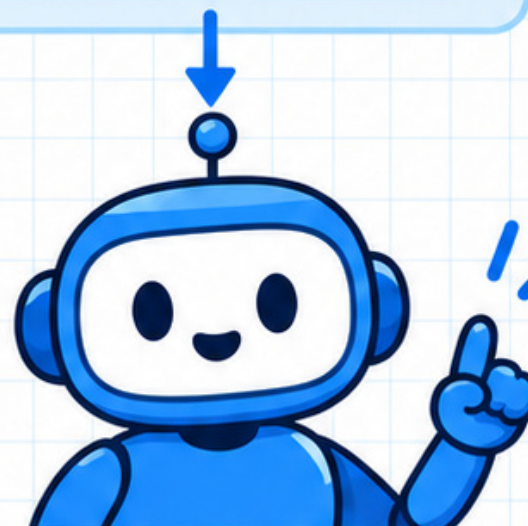
Policy

Determines eligibility.



Refund Policy

- Eligibility criteria
- Required documentation
- Terms and conditions



@learn.machinelearning

Use the photo,
the order record,
and the policy
together.

KEEP SENSITIVE CONTEXT SMALL

Toy support records

Include only authorized data needed for this task.

Remove secrets and unrelated customer details before the model call.

Permission checks happen before protected evidence enters the packet.

Application Data



Customer Profile

Name: Alex Chen
Email: alex@example.com
Plan: Pro



Support Ticket

Ticket: #4821
Issue: Cannot log in



Unrelated Record

Payment Card
4111 1111 1111 1111
(sensitive)

Permission Checks



Verify access to each record before it's included.

Assemble Context



Include only authorized, relevant data needed for this task.

Model Provider



Excluded
(unrelated or not authorized)

A smaller, cleaner context is safer and more focused.



DEFINE OVERFLOW BEHAVIOR

If the packet exceeds its budget: remove duplicates, select stronger evidence, or summarize allowed history.

Recount after each change.

If required facts still cannot fit, narrow the task or return a clear limitation.

1

Count the packet.

2

Remove duplicates or compress allowed content.

3

Recount and verify required facts.

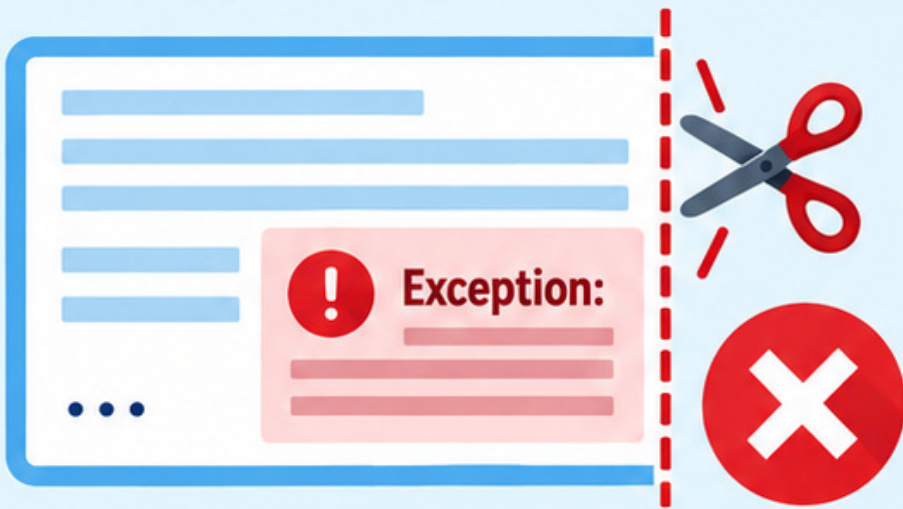
4

Within budget: proceed.
Still cannot fit: narrow the task or state the limitation.



DO NOT SILENTLY TRUNCATE

Cutting the end can remove the only exception.



Cutting the start can remove the task rules.



Choose context deliberately and verify required-fact retention after assembly.

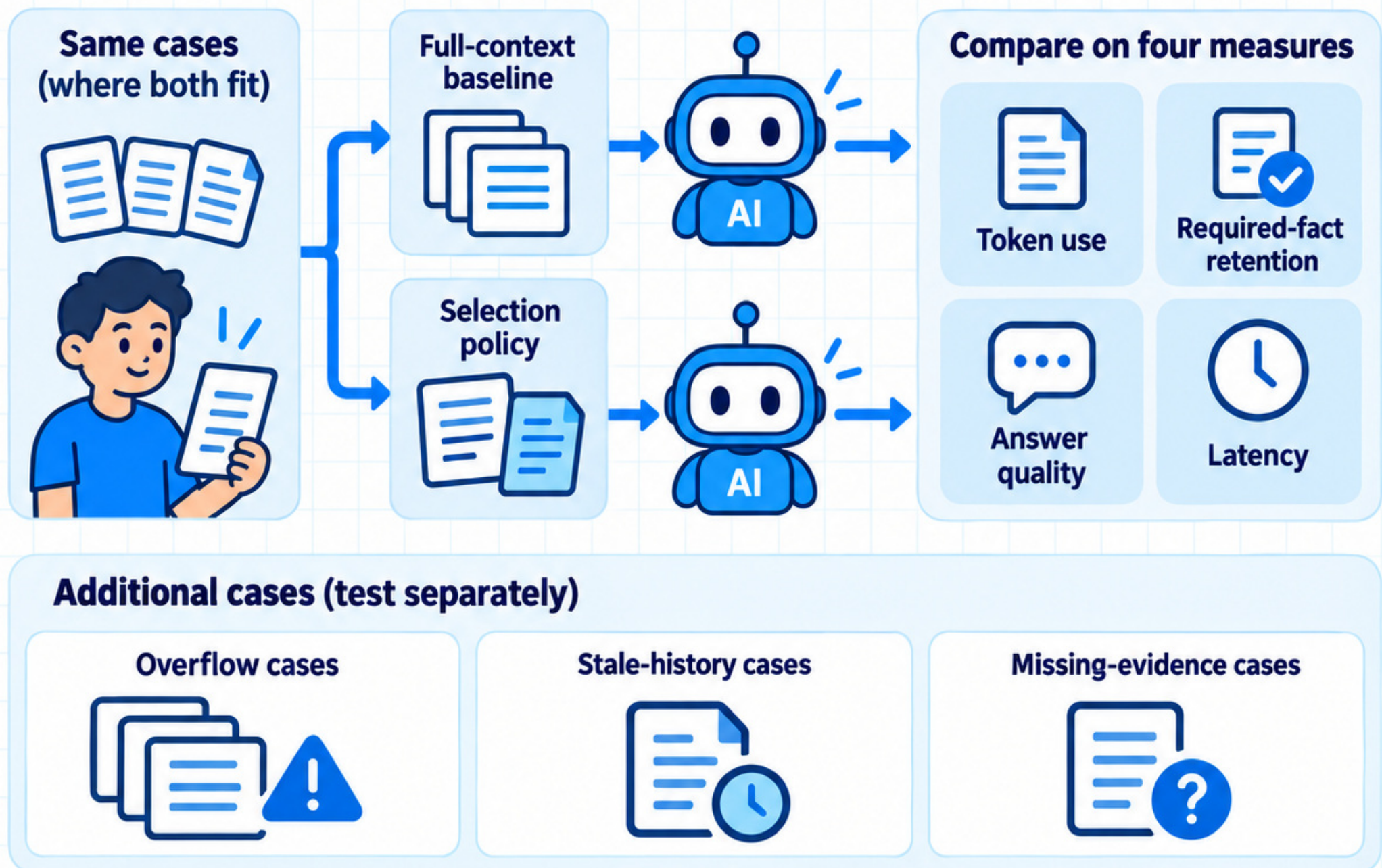


Selected context



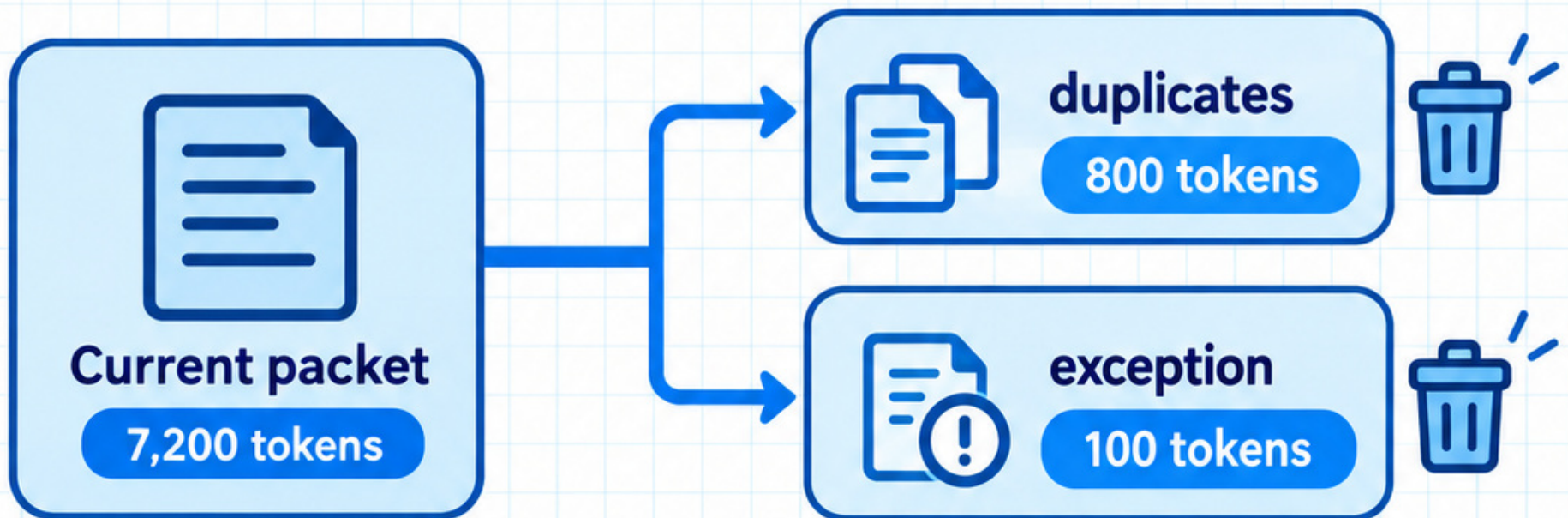
TEST THE CONTEXT POLICY

- Compare selected context with a full-context baseline on cases where both fit.
- Measure token use, required-fact retention, answer quality, and latency.
- Add overflow, stale-history, and missing-evidence cases.



TRY THE OVERFLOW CASE

Toy packet target: 6,500 tokens. Current packet: 7,200.
Duplicates use 800 tokens. The return exception uses 100.
Which should you remove first, and what is the new packet size?
Assume duplicate removal preserves all unique facts.

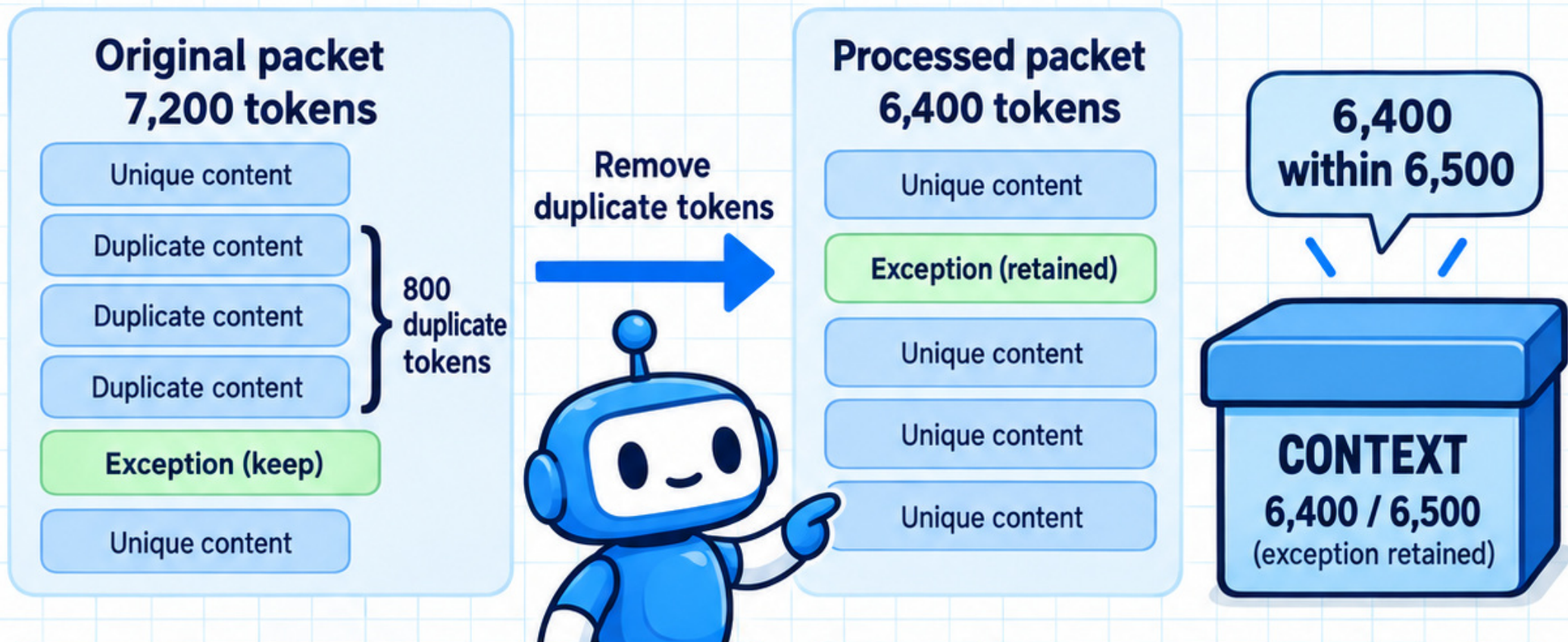


Which should you remove first,
and what is the new packet size?



BUDGET FOR USEFUL EVIDENCE

**Exercise: remove the 800 duplicate tokens.
The packet becomes 6,400. Keep the exception.**



A context policy needs limits, priorities, traceability, and failure behavior.



Next: generation and streaming.