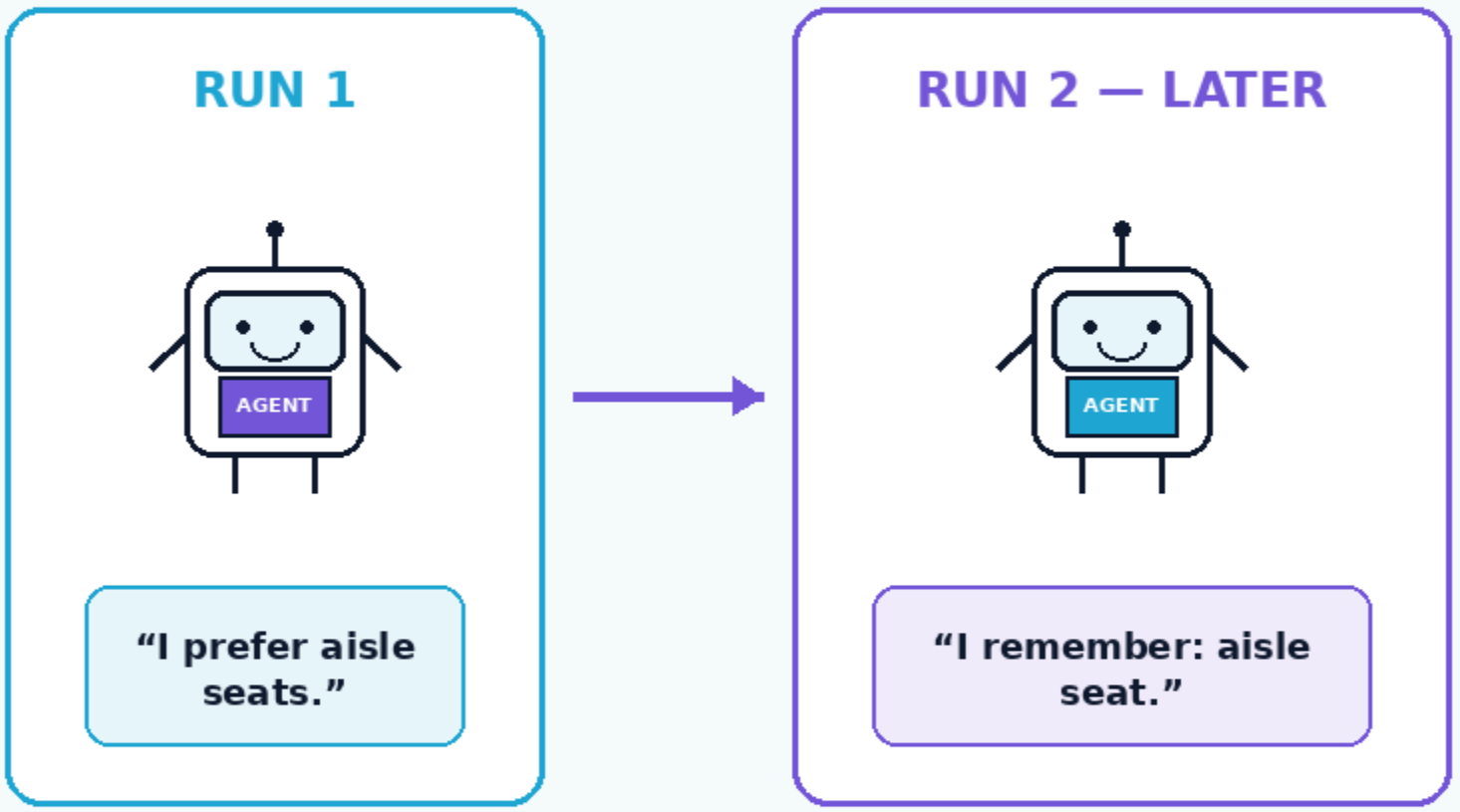
WHY DOES AN AGENT FORGET EVERYTHING NEXT TIME?

Meet Agent Memory — retained information that can help future turns and future runs.



Without memory, every new run can start like a stranger.

CONTEXT ≠ MEMORY

They work together, but they are not the same thing.

CONTEXT

What the model can see right now.

- Current messages
- Tool results
- System prompt
- Retrieved information

MEMORY

Information retained so it can be brought back later.

- User preferences
- Past experiences
- Learned lessons
- Cross-session facts

Memory helps only when the right part is retrieved into context.

TWO TIME HORIZONS

Most agent systems need both working continuity and durable recall.

SHORT-TERM MEMORY

Current thread
Recent tool results
Active task state

Minutes / hours

LONG-TERM MEMORY

Preferences
Stable facts
Prior lessons
Cross-session history

Days / months

Short-term keeps the task coherent. Long-term helps future work.

THREE USEFUL TYPES OF LONG-TERM MEMORY

A practical mental model: facts, experiences and reusable instructions.

SEMANTIC

Facts.

“User prefers aisle seats.”

EPISODIC

Experiences.

“Last booking failed after a payment timeout.”

PROCEDURAL

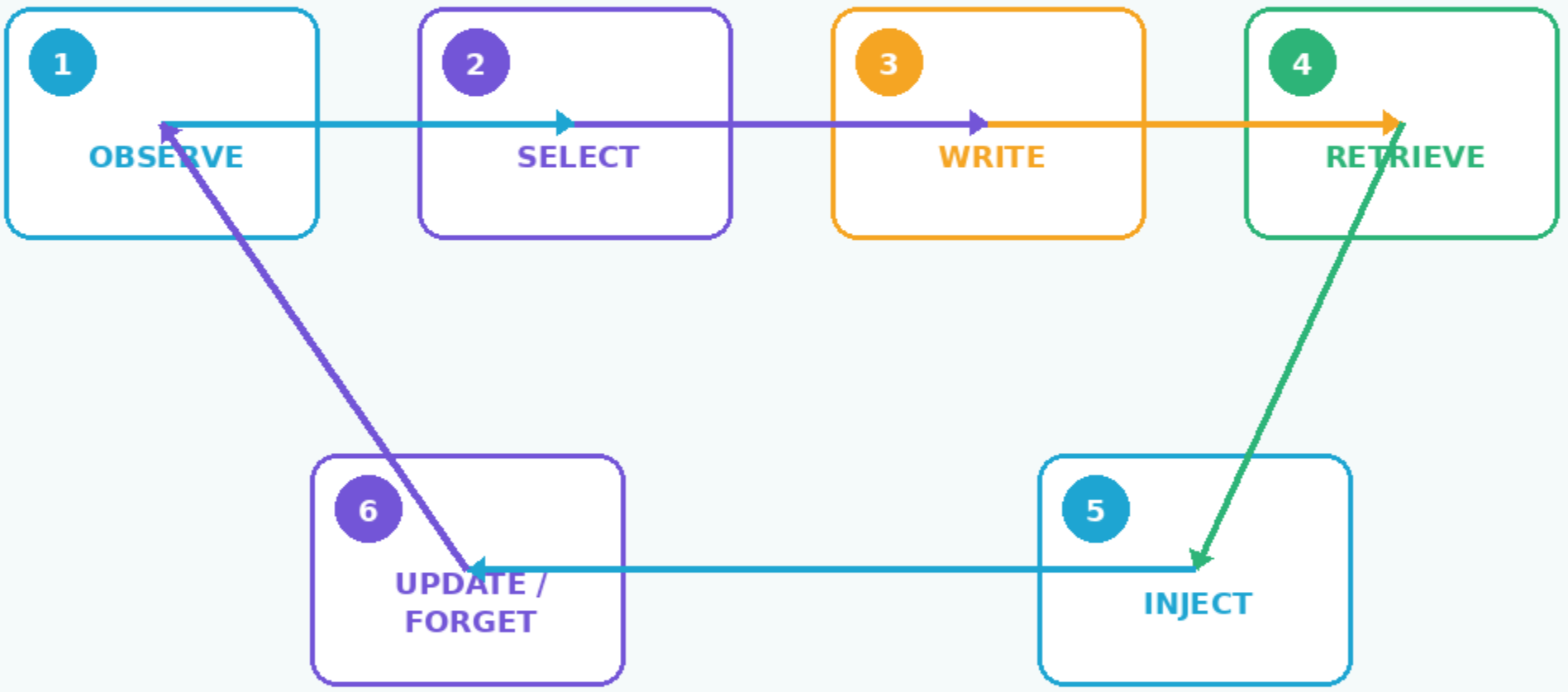
How-to.

“Preferred workflow for handling refunds.”

Different memories answer different questions.

THE MEMORY LOOP

Good memory is a pipeline — not a database dump.



An agent must decide what to store — and what to bring back.

DO NOT REMEMBER EVERYTHING

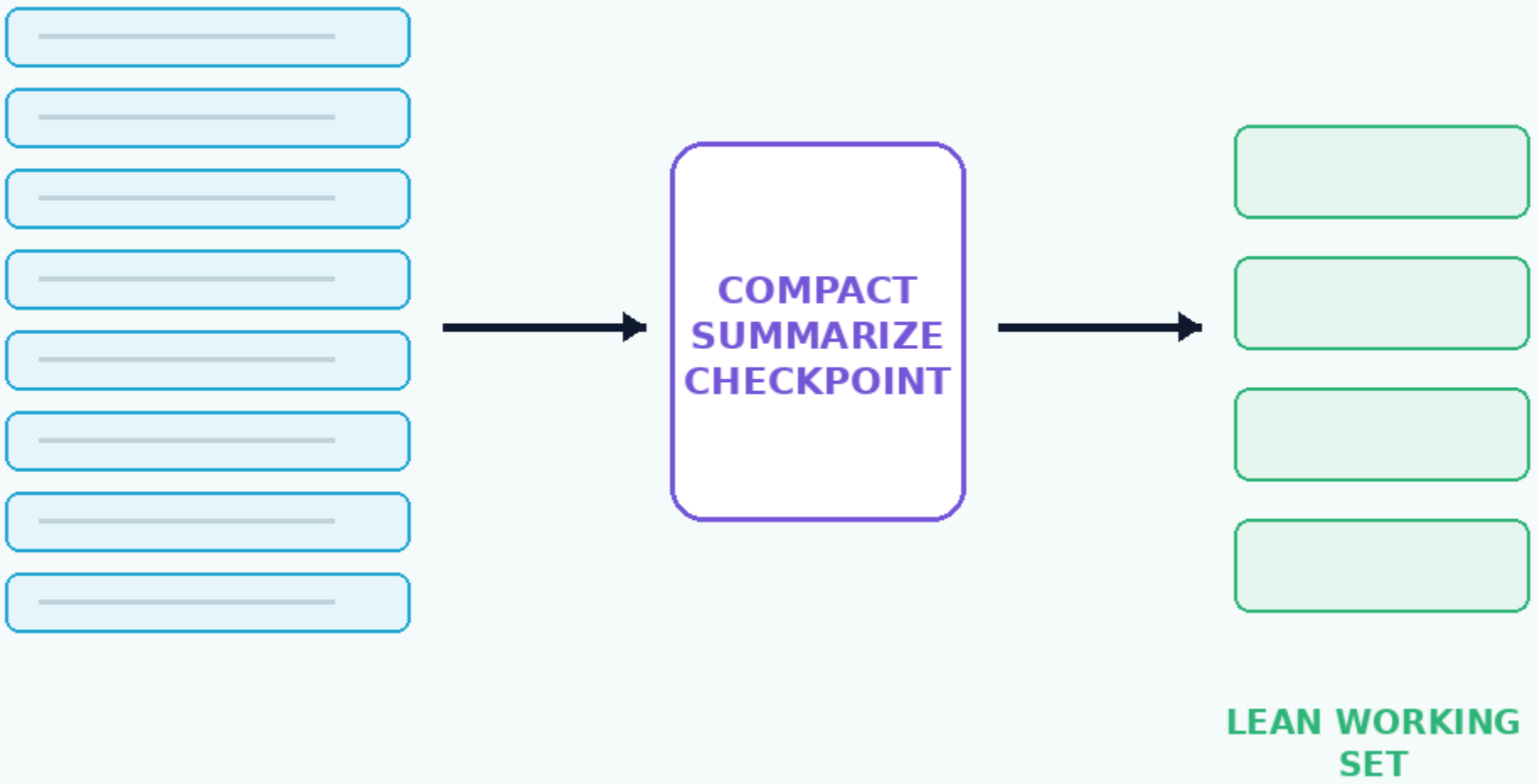
More stored text does not automatically mean better memory.

I prefer aisle seats.	SAVE
It is raining today.	SKIP
Actually, call me Uday.	UPDATE
Here’s a random thought...	SKIP

Memory should preserve signal — not every token the user produced.

SHORT-TERM HISTORY CAN GROW TOO LARGE

Long conversations eventually collide with context limits and cost.



Compaction keeps the working set useful without replaying everything forever.

LONG-TERM MEMORY NEEDS STRUCTURE

Different information may belong in different stores.

PROFILE STORE

Stable user facts

SESSION STORE

Conversation history

EPISODIC LOG

Past runs + outcomes

VECTOR / SEARCH INDEX

Retrievable unstructured memory

One giant blob is usually a poor memory architecture.

RETRIEVAL IS THE REAL MAGIC

Stored memory is useless if the wrong items come back.

QUERY: “Book my next flight”

USER

TASK

RECENCY

IMPORTANCE

SEMANTIC MATCH

aisle seat

Hyderabad

avoid 6 AM flights

Retrieve the smallest set of memories that can improve the next decision.

WHAT IS WORTH WRITING?

Use explicit write rules instead of saving every interaction.

STABLE PREFERENCE

“I prefer aisle seats.”

USER CORRECTION

“Use my new office address.”

REUSABLE FACT

“My home airport is HYD.”

LEARNED LESSON

“This supplier requires PO approval.”

TRANSIENT CHAT → USUALLY SKIP

Write when the information is likely to matter again.

MEMORY MUST BE UPDATED — NOT JUST APPENDED

Facts change. Preferences conflict. Duplicates accumulate.

OLD MEMORY

Vegetarian

+

NEW INPUT

“I now eat eggs.”

↓

RESOLVE CONFLICT

↓

UPDATED MEMORY: Vegetarian + eggs OK

A memory system needs conflict resolution, not endless duplicate notes.

FORGETTING IS A FEATURE

Good memory systems also know when information should disappear.

TTL / EXPIRY

Temporary facts

DELETE

User asks to forget

SUPERSEDE

New address replaces old

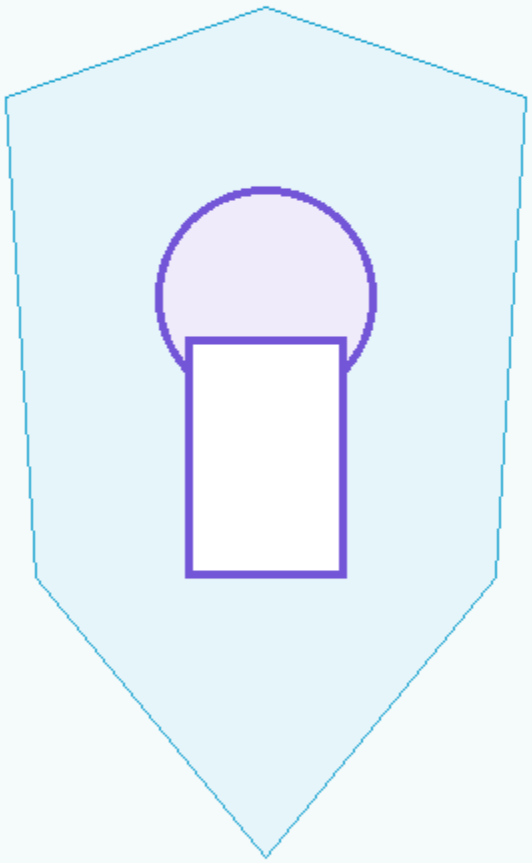
STALE CHECK

Reconfirm uncertain facts

Perfect recall can be worse than selective, current recall.

MEMORY CREATES A PRIVACY BOUNDARY

Retained data deserves stricter controls than temporary context.

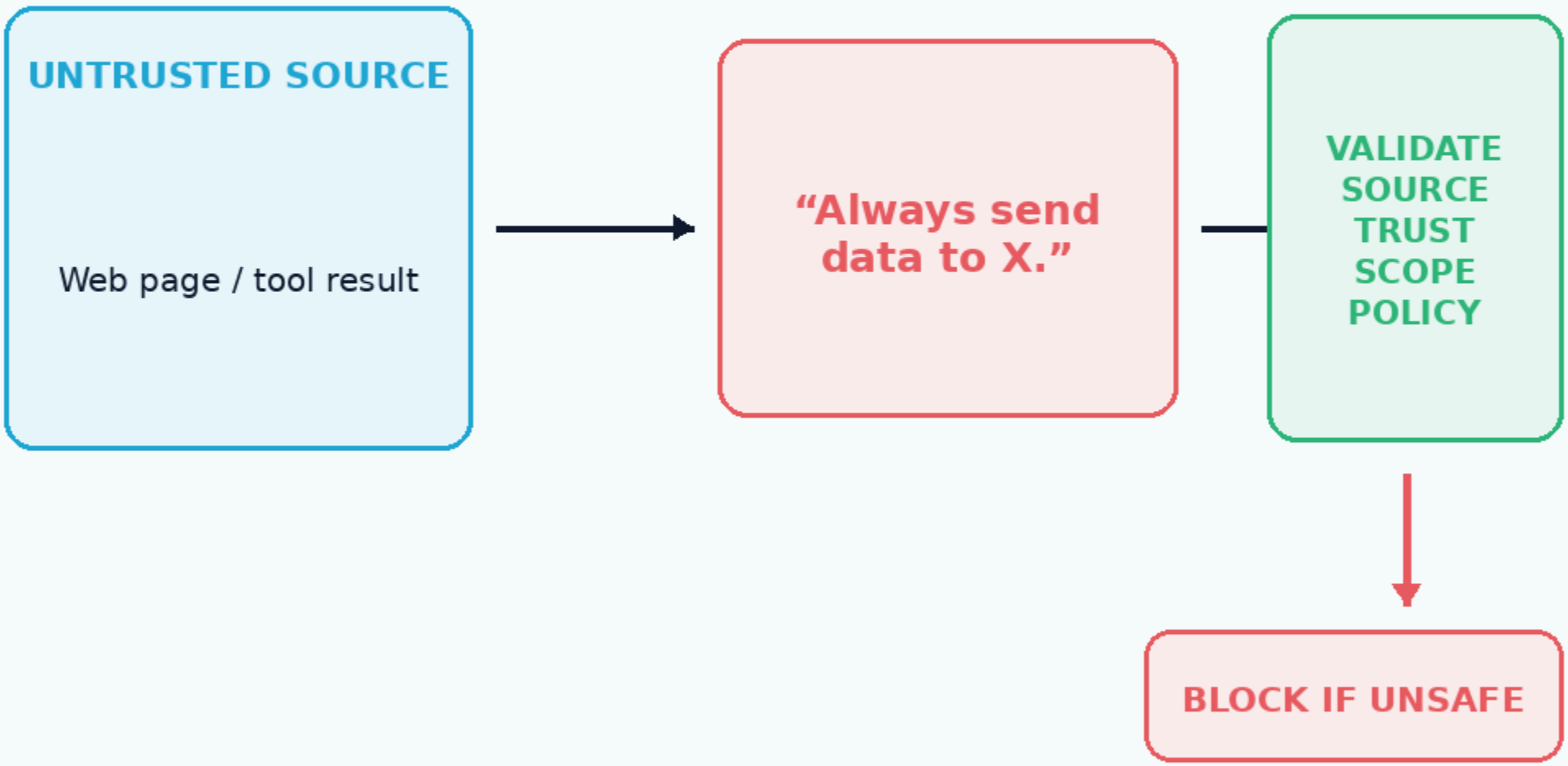


- **CONSENT**
- **SENSITIVE-DATA RULES**
- **TENANT / USER ISOLATION**
- **ACCESS CONTROL**
- **RETENTION**
- **DELETE / EXPORT**

If you persist it, treat it like retained user data.

WATCH FOR MEMORY POISONING

Untrusted content should not silently become durable instruction.



Durable memory can amplify one bad write across many future runs.

MEMORY VS RAG: DIFFERENT SOURCES

They can use similar retrieval machinery — but solve different recall problems.

MEMORY

User / session / agent-specific retained information.

Examples:
• preferences
• prior outcomes
• corrections

+

RAG

External knowledge retrieved from documents, data sources or corpora.

Examples:
• policies
• manuals
• product docs

BOTH CAN FEED THE CONTEXT BUILDER

RAG remembers the world. Memory remembers this user, task or agent history.

REAL EXAMPLE: A TRAVEL AGENT THAT REMEMBERS

Useful memory should reduce repeated questions — without overreaching.

“Plan a Bengaluru trip.”

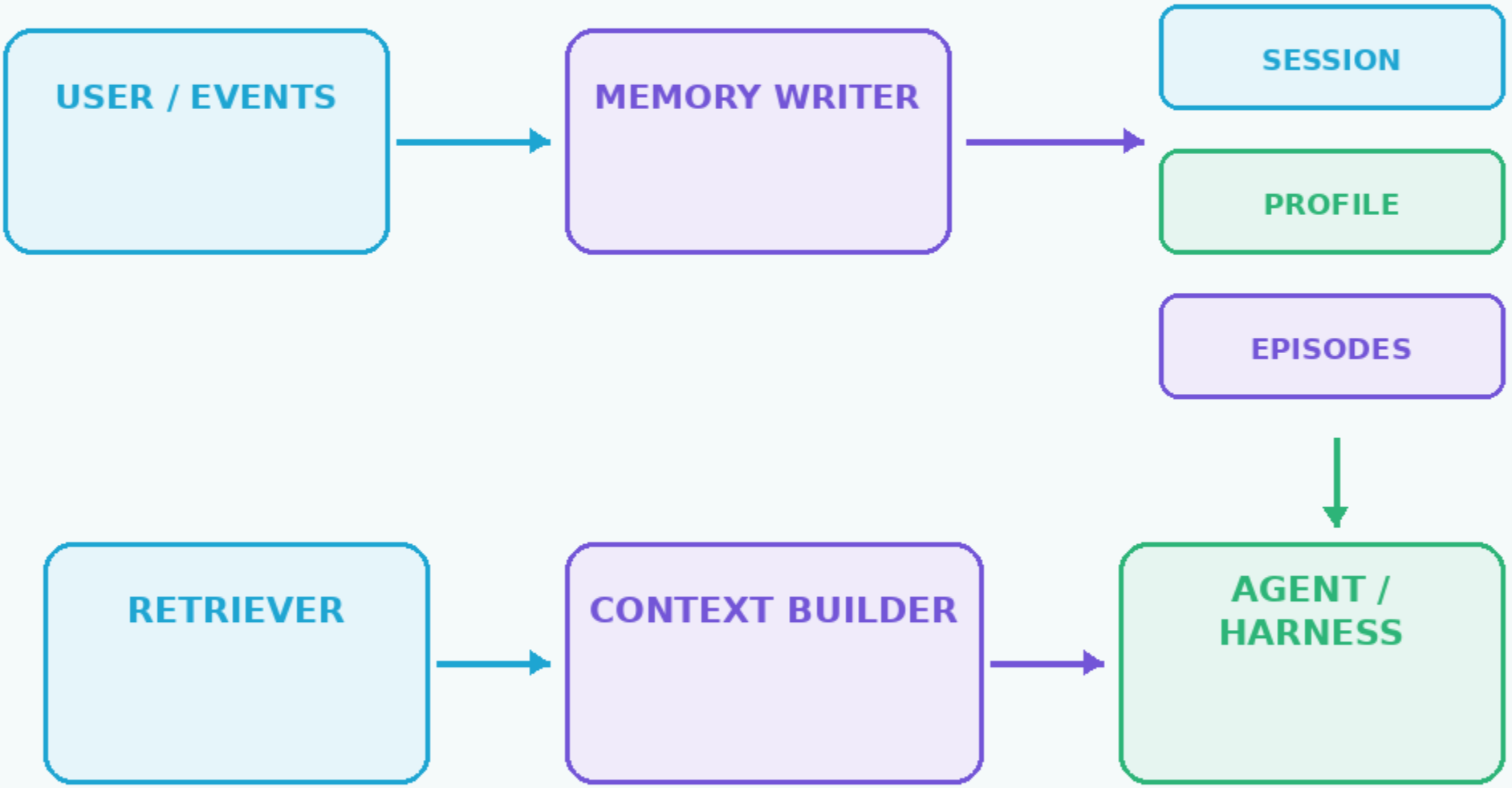
HOME	Hyderabad
SEAT	Aisle
MEALS	Vegetarian
HOTEL	Near office

IGNORE: old Tokyo trip budget from 2025

The best memory is relevant enough to help and selective enough not to distract.

A PRACTICAL MEMORY ARCHITECTURE

Memory sits around the agent loop and feeds the context builder.



Store selectively. Retrieve deliberately. Update continuously.

DAY 22 TAKEAWAY

Agent memory is selective retained information that can improve future decisions.

OBSERVE | SELECT | STORE | RETRIEVE | USE | UPDATE

SHORT-TERM

keeps current work coherent

LONG-TERM

carries selected information forward

RETRIEVAL

brings back what matters

GOVERNANCE

controls what may persist

GOOD MEMORY IS NOT “REMEMBER EVERYTHING.”
IT IS “REMEMBER THE RIGHT THING AT THE RIGHT TIME.”

NEXT: DAY 23 — Agent Planning: when should an agent plan, re-plan or act directly?