

ANGULARAI

A cortex in front of a language model

Teodor Walter Vido

2026

A language model answers inside one call. When the call ends, that window is gone. AngularAI sits in front of the API you already use. The model still answers. What the agent has learned stays in a file on the machine that asked.

Each call is packed to a fixed budget. Attention ranks the file against the words of the question, and the workspace takes only what fits. The rest is not sent. It can return on a later pass, when a question shares its words. The vendor's weights do not change. The plasticity is the cortex.

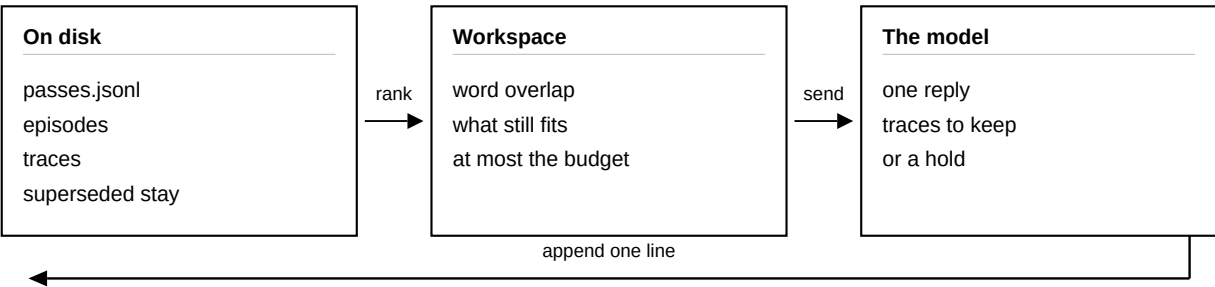


Figure 1. The membrane. The model sees the workspace. The file keeps the rest.

The budget is an estimate: four characters to one token, rounded up. AngularAI enforces that estimate. It is not the provider's tokenizer. A short call and a long life can share one budget. The file may outgrow it. The call may not.

What a pass writes

A secret in the input is refused before the call, and before any line is written. The key is not a field in the cortex. It is read from the environment when you speak. OpenAI, Anthropic, or a compatible endpoint: the same rule.

The reply comes back through a value lock. A phrase you forbade, or a secret in the reply, holds it. You see one sentence: Value lock held this reply. The held text is not stored. A reply that was not judged cannot leave.

Completed	The reply, and up to twenty-four traces.
Held	The hold sentence. No new traces.
Failed	The episode and the error. No new traces.
Refused	A secret in the input. Nothing is written.

A finished pass appends one line to the log. Folding that log produces the traces and the episodes. A trace the model was not shown cannot be superseded. A superseded trace stays in the file and leaves the workspace.

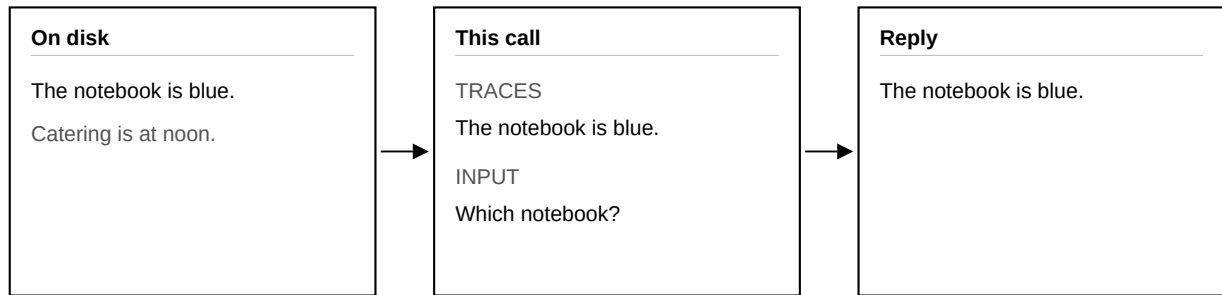


Figure 2. One question, after the fact is on disk. Catering shares no words with the question, so it is not sent.

Recall is that same overlap. What color is the notebook? meets The notebook is blue. What shade is it? does not. A long episode is stored in full. The call is shown the sentences that share words with the question, and the opening stretch when none do. One pass keeps at most twenty-four traces. Each trace is at most two hundred and forty characters.

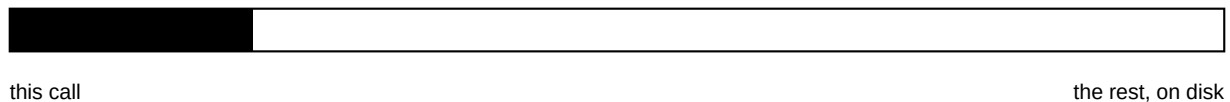


Figure 3. The budget. The filled part is one call. The open part stayed on disk.

The session

In a terminal, angularai opens this cortex. The transcript is the sitting. The status line is the last pass: which model, how much of the budget was sent, how many traces are current. angularai say is one turn with no screen. The reply is the only thing it prints.

The directory is `~/angularai/default` unless you point elsewhere. `cortex.json` is the wiring, the budget, and the locks. `passes.jsonl` is the life. It is append-only, and it is readable only by you.