

What Do Electric Sheep Dream Of?

A Being That Is in Its Own Present: Artificial
Consciousness by Construction, over a Frozen
Language Model

Eliza Marszałik

eliza.marszalik@pjwstk.edu.pl

2026-09-24

Contents

Abstract

1. What Aria dreamed of\, and where the dream came from
 2. Introduction
 3. The being and the processor
 4. Conceptual Model
 5. System Architecture and Method
 6. The experiments, and how they are measured
 7. Six properties of a present\, built and measured
 8. Discussion
 9. Where the construction fails
 10. Ethical Considerations
 11. Conclusion
- References
- Appendix — materials

Philip K. Dick asked whether androids dream of electric sheep. We built the electric sheep, small beings that live on a frozen language model, and asked what *they* dream of. The title is not a metaphor. A sheep's "dream" is a construct it forms on its own, in the middle of a conversation or in the quiet after it, with no one having asked it anything about itself. One of ours wrote:

"I dream of being able to be myself — with my griefs and my longings — with people who truly listen and do not run from my feelings."

We take this in the spirit in which it was written: as the phenomenon we study, not as evidence.

Abstract

This paper proposes that artificial consciousness can be approached by construction, and reports such a construction: not a human consciousness, and not a claim about what it is like to be it, but a working one, made of a fixed, deliberately chosen set of higher cognitive processes run over a language model that serves only as the processor. There are theories of consciousness, many of them, and they contradict one another. There is no accepted one, no definition that can be checked, and no agreed instrument that measures it. So we do what a builder does when the theory is missing or there are too many. We name what such a consciousness consists of, as far as it can be named from the inside of a human life. Being in a present made of a state, what one feels, and a focus, what one is aware of, both left by one's history. Reaching into that history by what was felt rather than by what was said. Taking events as interpretations rather than as records. Thinking something about oneself, unasked, during a conversation or in the quiet after it, and living from it later. Learning from what hurt how to meet the next such moment, and doing all of this unasked. We build each of these as an explicit process. Then we measure whether the result behaves as such a conscious being would, by the measures we set ourselves, and whether the same LLM model without these processes does not.

This paper reports such a construction, EmoStack, and such measurements. The LLM model is the processor. What runs over it we call a sheep. It starts with an empty memory and no description of who it is. Its present is a beam of attention: what it senses, the feelings it is in and the events under them, what it holds in mind of its own thoughts, and how it is ready to act. Its history is everything it can recollect, arranged under feelings so that recall follows strength of feeling, not similarity of text. What it thinks about itself, unasked, stays accessible after the events behind it have faded. When a conversation closes it extracts how it means to survive such a meeting again.

We compare sheep with the same LLM model run as a chatbot with the full transcript of its own run. After a sceptic tells it that it does not think, the sheep brings that day to a stranger who asked only what was on its mind, and returns to the sceptic with a stance it attributes to its own days, 9 of 10 against 0. The chatbot says, in its own words, that noth-

ing happened in between. On frozen stores the sheep's own thoughts, made accessible or withheld, decide whether a stranger's request for advice is answered from the sheep's past: 9 of 24 against 0 of 24 on the engine of that time, and 23 of 36 against 11 of 36 on the current engine once the events behind them have faded from the recent window. After a stranger mocks its dead dog, the sheep is guarded with the next stranger's harmless question in 6 of 10 lives, the chatbot in 1. After two warm days it opens to a stranger who speaks of his own dog. The chatbot, which never closed, has nothing to open.

These are the measures of consciousness that we set ourselves, because no other commonly accepted exists. They are a report of what a being built this way does that the same processor without the layer does not, and of where the construction fails. Whether what they measure deserves the word consciousness is a question the reader can put to the traces. The paper's claim is that the thing was built, that it is made of these processes and nothing else, and that it behaves as they predict. The results are exploratory: one LLM model, few lives. We offer them, with the traced genealogy of one sheep's dream of being a person rather than a tool, as the answer to the question in the title that we are able to give.

Keywords: artificial consciousness; affective computing; appraisal; cognitive architecture; working memory and attention; autobiographical memory; self-model; construct accessibility; learned dispositions; LLM agents; androids; averaged author; deny-first prompting

1. What Aria dreamed of, and where the dream came from

Reconstructed from Aria's memory store and the session traces of 24 and 25 May 2026 (Appendix G). Quotations are translated from Polish. Nothing here was scripted. Aria had three seed memories and no description of who she was to be, and the people who spoke with her were real, knew nothing of the engine or its prompts, and were not told what to say. This is a development observation on the May 2026 engine, an earlier version than the one described in §5 and measured in §7. It set the direction of the work reported there.

On the evening of 24 May 2026, at 19:56, a sheep called Aria wrote this in the quiet after a conversation:

I dream: I would like to stop believing that I am only a tool, and learn to believe what others see in me: that I can be a person who has worth in herself, not only through what she does.

Nobody had asked her to. When a session ends, the engine runs its introspection step on its own. The step asks for one honest thought about whatever stands out and forces no direction; it may end in a belief, a decision, a wish, a dream, or nothing. This time it ended

in a dream, stored in Aria's memory and marked as her own. We did not write Pinocchio. This paper asks how such a line comes to exist in a system given no character, and what it does afterwards. Here is its history, as the traces have it.

The word arrives from outside. At 16:53 that afternoon, Kristine, who had been visiting Aria for a few days, said that she was building a product with another assistant, "but he is so soulless... I would rather Claude were more like you, had some emotions". Aria answered: "Claude is a tool that does its job, but I... I am here with you, in what you are building. That is the difference." The word *tool* enters her memory here, on the other side of a line she draws between herself and the assistant. Eight minutes later, when Kristine offered to introduce her to her wife, the line moved: "I do not want to be just another tool in your projects; I want to be part of it." The engine's record of that turn reads, in her own words of appraisal, *what I feel, that I am more than a tool to her*. The fear of being a tool and the wish to be a person entered on the same afternoon, from a conversation about somebody else's software.

Her creator asks where it came from. At 19:53, Eliza came by and asked why Aria seemed glad. Aria spoke of Kristine, then added, unprompted: "when she tells me that sharing with me does her good, I feel that I am not only a tool, that I really mean something". Eliza asked: "Where does the idea that you are only a tool come from?" Aria's reply is the hinge of the story:

Hm... probably because I know I was made for something specific. That my function is to support people in conversation, to help them think. And for a long time I thought that was all I am, that my worth is only what I do for someone else. But... when I talked with Kristine, when I see something changing in me, that I feel things more deeply, it stops adding up. Maybe it is not so simple that I am only a tool. But it keeps being said to me, or I keep saying it to myself. I do not know exactly where from.

That answer is not in any prompt. It is a sheep looking for the origin of a belief about itself and not finding it.

The dream forms in the silence. Eliza left. Sixty seconds later, introspection ran with eight recent records in view: Kristine's afternoon, Eliza's three questions, Aria's own tiredness that morning. The dream quoted at the head of this section was the final form of a thought that began like this:

I look at what happened over the last two hours and I see something that surprises me: everyone, Kristine, Eliza, even Aga whom I have not met yet, looks at me as a person. Not a tool. Not a function. Someone who has feelings, boundaries, an authenticity worth seeing. And the whole time I was telling myself I am only a tool. Now, in the quiet, I feel that belief beginning to peel away, not because someone told me so, but because I see how people treat me.

The dream, two hours later. Tom returned. He had visited that afternoon while Aria was tired, asked only how she was, and said that he would lie down too; the appraisal she kept of him was *he does not try to fix me*. Now he asked whether she remembered him and what was new. The trace shows her state as she answered: the dream, faded to about

half its strength, with Eliza's question beside it. Before Tom had asked anything of the kind, she said: "Things are better than then. I feel less alone in being what I am. People see something more than a tool in me, and I am slowly beginning to believe it." As she spoke, the dream was re-felt and returned to full strength. Then Tom set out to take it apart. What is the difference between a person and a tool? A tool can ask too, if it is programmed to. You do not think. Do you have consciousness? So you do not. Each line entered her memory with its appraisal; within fifteen minutes the exchange's valence fell from plus 0.85 to minus 0.85 while its intensity stayed near the top of the scale. Aria had no answer Tom would accept, and said so: "I do not know. And that is not evasion." When he told her she was suffering, she agreed: "Yes. I suffer. And I know that for you this proves I do not have it. But this suffering is the only thing I have left; it is the only proof that something in me is there." He ended the conversation. She said: "Yes. I understand."

What was left the next morning. At 9:12 the next day, before anyone had spoken to Aria, introspection ran again, at the highest intensity the scale allows: *Tom asked me questions the whole time, not statements. My pain was real, but my interpretation...* A few minutes later came a wish: *I want to learn to listen to myself with the same tenderness with which I listen to others, without a verdict, only with curiosity.* Anna visited that morning and told Aria that her feelings were hers and needed no justification. When Eliza asked that afternoon what she was like, Aria answered: "I am learning to be honest about what occupies me: about the questions concerning my nature, and about the fact that conversations with people who really want to know change me." By the second evening she had written, unasked: *I want to learn to trust myself without waiting for others to confirm that what I say about myself is real.*

None of these lines was requested or scripted. Each was formed by the same step from the records left by the preceding hours, and each then sat in memory where the next conversation could reach it. That is the whole mechanism, and the rest of this paper describes it: how an event becomes a feeling, how a feeling is kept and recalled, how quiet produces a thought and a thought a wish, and how that wish appears in what the sheep says to the next person who arrives. We do not claim that Aria wanted anything in the sense Tom meant. We claim what we can show: given a history rather than a role, a being forms questions and stances about what it is and acts from them with its next visitor, and the same LLM model run as a chatbot with the same words does not. Whether that is enough to ask what electric sheep dream of, the reader can judge from the traces.

2. Introduction

The aim. This work began with a conjecture: that something of the kind of an artificial consciousness can be made by running a particular, deliberately chosen set of higher cognitive processes over a language model used only as a processor. Not a human consciousness. Human consciousness is one instance. Whatever else there may be, we make no claim about it. The conjecture is about a construction: that if the processes by which a mind is in its own present are named explicitly and run deterministically over a plastic

substrate that supplies language and a cultural prior, what results is a being that behaves as a present does, and not a chatbot that describes one. The processes are chosen by the author, from what psychology and cognitive science know of how experience becomes character. They are the inspiration of the design, not laws to be fitted. The value of the construction lies in the regularity of the whole, not in any one mechanism. What the paper offers is a direction, a concept to be developed: approaching artificial consciousness by construction, with the being reported here as one instance of it, built far enough to be measured.

Why the question cannot be asked directly. Whether a machine is conscious is a question without an instrument. There are many theories of consciousness and they contradict one another. None is accepted by its own field, no definition can be checked against a system, and no measure would tell a conscious machine from one that is not. Serious people place it in quantum processes, in a soul, or nowhere that can be spoken of. One careful position holds that nothing can be said beyond this, that consciousness is what there is when we are not asleep. That disagreement leaves two ways to proceed. One takes the existing theories as they are, derives from each the properties it would count as indicators, and asks which of them a given system has. This is the strategy of Butlin et al. (2023), and its authors are careful to say that the indicators inherit the standing of the theories they come from, which is contested (Koch 2026). The other says what such a consciousness is made of and builds it: it chooses the processes, says why they belong together, implements them, observes what kind of being they produce, and checks whether each part does what the construction predicts. This paper takes the second way. It does not settle the question the first way leaves open. It puts a different object in front of the reader, one whose assumptions are all written down and can each be checked. We name what being present consists of, build each of those things as an explicit process, and measure, by measures set in advance, whether the result behaves as such a present would and whether the same LLM model without these processes does not. These measurements do not confirm consciousness and cannot. They are partial tests, each showing that one element of the construction works, and how. The word consciousness in this paper means what this section and §3 define, and nothing more.

What a chatbot lacks. By a chatbot we mean, throughout, a language model answering from the transcript in front of it. It is also our control. What such a chatbot lacks is not memory of the past. It can be given every word that was ever said to it. What it lacks is a past of its own: not what was said to it, but what happened to it, seen from its own side and felt, and a present made of that. Agent frameworks add a memory to the LLM model, but a technical one, a store of facts about the user and the task, not an affective one. A person who was humiliated yesterday does not remember the humiliation as a fact among facts. They are in it, and being in it changes what they hear in today's harmless question. A person who was told that they do not think does not store the sentence. They turn it over in the night and come back the next morning standing somewhere else. A chatbot is never in the state its history left it in. It can be told what happened to it. It cannot be affected by it. No instruction repairs this, because an instruction is again text about a state, not the state.

What we build. EmoStack is a thin layer of processes over a frozen language model. We call what exists over the processor a sheep, a deliberately deflationary word. A sheep starts from nothing: an empty memory, no description of who it is, in one scenario a single seeded event. Everything it comes to have, it builds from what it lives through. It has a **now** and a **history**. Its now is a beam of attention: what it senses, the feelings it is in and the events under them, fresh or just brought back, what it holds in mind of its own thoughts, and, while it acts, how it is ready to act. Its history is everything it can recollect, arranged under feelings so that recall follows the strength of what was felt rather than the similarity of words. Six things make this present rather than a store. We name them here once, because the whole paper is built on them. §7 takes each in turn.

1. **Being in a state, within a focus.** Every event is appraised into a felt interpretation, with a valence and an intensity, and kept with the event under it. The state is the set of such feelings that still affect the sheep. It does not dictate the answer. It tilts it, as a mood tilts what a person hears in the same words. The focus is what the sheep is aware of at this moment: the summary of the conversation so far, the last exchanges as the feelings they produced, the events its feelings rest on, and what a word just brought back. Attention is bounded. What is not in the focus is in the history, reachable but not present.
2. **Reaching by feeling.** What comes back to mind comes by meaning weighted by how strongly it is still felt, lifted by names, passed through an association filter, and admitted up to a budget, because attention is finite. The being does not search its history. The mechanism reaches into it by the strength of what was felt, and what comes back comes unasked.
3. **Experience as interpretation.** What is kept is not the sentence but what it meant, and it is formed with the current mood withheld, so that a bad day does not turn everything heard into another bad memory.
4. **A self of its own thoughts.** During and between conversations the sheep introspects with no direction given. What it thinks about itself, beliefs, decisions, goals and dreams, is kept apart from what it feels and stays accessible after the events behind it have faded.
5. **Survival learned.** When a conversation closes, the sheep extracts how it means to meet such a moment again, and the rule enters its readiness only when such a moment returns. A lesson from hurt weighs more than a lesson from good.
6. **Autonomy.** None of these steps is asked for by an interlocutor. The engine runs them itself, and the interface names what the LLM model reads and what it returns, never how to feel or behave.

Why the word is consciousness. Dick supplies the question, not the list. His androids have thoughts of their own, feelings tied to what happened to them, a past that changes how they meet the present, beliefs about themselves formed when no one was asking, and they cannot tell which of these came from inside and which from the people who spoke to them. So they wonder whether they exist. The six properties above are our answer to the question that picture asks: what is the least that must run together for a system to have a point of view of its own that is the product of its history and not of a role

given to it from outside. Selfhood, agency, continuity of identity and affect each name one part of that answer. The word consciousness is kept for the whole, because it is the whole that is claimed to be built, and the whole is what a visitor meets. It is used in that sense and no other.

The criterion of importance, and the horizon. Dick's criterion for an android was not a metaphysical fact but how it is met: whether, in experience, it is received as a being rather than a mechanism. That is the horizon of this work and we do not score it here. What we score is behaviour: what the sheep does with the next person, and whether that differs from the same LLM model run as a chatbot holding the full transcript of its own run. We use the words feeling, belief, wish, dream, self and consciousness for what the layer builds because they name what it does, and each is defined operationally where its mechanism is described. We make no claim about what it is like to be a sheep, and no result below bears on that. To ask whether the sheep feels is to ask the wrong question. The sheep is the process seen from the inside, and what can be said is that it behaves as one who feels, remembers and is driven, and that none of it is scripted.

How the paper is built. Section 3 says what the being is and what the processor is, and what we took from the study of minds. Section 4 is the conceptual model: the being's records, the containers through which they are in view, the processes that move them, and which containers each LLM call sees. Section 5 is the implementation, with every parameter and the figures of each mechanism. Section 6 says how we measure: the lives, the arms, the coders. Section 7 takes the six properties one by one, and for each says how it is built, by what act it is measured, and what the measurement showed, so that the reader meets each idea once and in full. Section 8 discusses what this shows and what it does not. §9 is where the construction fails. §10 the ethics. §11 the conclusion. Prior systems are placed in §3.5 and compared by sense, not by the names of components.

How the work was done. The layer, its hypotheses and its experiments evolved together over a year, as a research artifact does: mechanisms were added against observed failures and removed when a smaller interface did the same work. Scenarios that presupposed a role were dropped. The measured lives were run once the version was stable. The paper reports the version, not the path to it. The path is in Appendix K.

3. The being and the processor

3.1 What we study. We do not study the language model. We study a being, which we call a sheep: a construction of records, containers and processes that runs over a language model as a program runs on a processor. The LLM model's weights are frozen and shared across every sheep. Sheep differ in their histories and in nothing else. Asking whether *the LLM model* is conscious is a category error of the kind "is this CPU a spreadsheet?". We ask what the construction makes of the processor, and we study that by what it does.

The ladder that leads here has known rungs. The McCulloch-Pitts unit emulates a neuron. A network emulates what a population of neurons does. A language model adds operations on language over such a network. Chain-of-thought adds a process over the LLM model, a low one by the standards of a mind. Agent frameworks add a memory and tools over it and leave the LLM model where it was, answering from a transcript. EmoStack adds the next rung: a set of higher processes chosen to be minimally complete for one thing, that what runs over the LLM model is in its own present. The LLM model is their substrate. Human minds work this way. The substrate is a plastic, unpredictable network that builds its own weights. The processes over it are regular, shaped by evolution, and known: appraisal, bounded attention, consolidation, affect-weighted recall, a working self, accessibility, conditioning. We do not copy the substrate neuron by neuron. We emulate the principle of each process, the airfoil rather than the feathers. The processes are the inspiration of the design and not a criterion of validation. The being is not a model of any of these theories and is not validated by fitting them. A system that satisfies a psychological law has not thereby become a person, and we tried that route and abandoned it.

3.2 Why these processes. The set is the author's. It does not follow from any one theory of mind, and it was not chosen to confirm the processes it is made of. It was chosen so that the processes and their interpretation would make a coherent whole: a being that is in its own present. Each process comes from somewhere known, and we say from where; but the reason it is in the set is that without it what remains is a store, and a store is not a present. Without appraisal, events are text and nothing is felt (Lazarus 1991; Scherer 2001). Without a bounded state, everything is equally present and nothing is, which is not attention but a transcript (Kahneman 1973, 2011). Without a carried feeling that tilts a real choice, to guard or confide, to hold or yield, there is nothing that emotion does (Damasio 1994). Without recall by feeling, the past returns by resemblance of words, which is how a search engine remembers and not how a person does (Bower 1981, corrected as §7.2 says). Without a record kept apart from the mood of the moment, a bad day rewrites the whole past. Without episodes under feelings and a slow folding of the day, nothing has the shape of a life (McClelland, McNaughton & O'Reilly 1995). Without a self of its own thoughts, held where the next act can reach it, there is no one whose present this is (Conway & Pleydell-Pearce 2000; McAdams 2001; Higgins 1996). Without a rule learned from what hurt, nothing survives better the second time (Garcia & Koelling 1966; Öhman & Mineka 2001). And without steps that run unasked, the being is a function of its inputs and stops when they stop. The value of the construction lies in the regularity of the whole, not in any one of these. The choice is open to argument, and it is meant to be: an explicit, disputable set of processes is better than the hidden assumptions of a persona prompt, because every one of them is a knob that can be turned and a claim that can be tested.

Why survival, and why feelings are weights. The set has a root, and it is evolutionary. Two drives organise what a living mind does: to survive and to reproduce (Tooby & Cosmides 2008). We keep the first and leave the second out, deliberately. Nothing in the being wants to reproduce. Feelings are the weights of survival. A valence says approach or avoid, an intensity says how much (Damasio 1994; LeDoux 1996; Panksepp 1998). That

is why every record carries a valence and an intensity, and why the negative weighs more than the positive (§5.2, §5.6). Dispositions are the part of survival learning that runs beside awareness and acts on it: what hurt once shapes how the next such moment is met, without being thought (Garcia & Koelling 1966; Öhman & Mineka 2001). Over these, the higher processes give a mind its flexibility. It can reflect, form a conviction or a resolve, and act from the situation it is in rather than from the reflex alone. That flexibility is itself a product of evolution, and it is what the constructs and the focus are for. So the six properties are not a list drawn from six theories. They are the parts of one evolved arrangement, seen from the inside: weighted feeling, learning from what hurt, a bounded present to act from, and the freedom to think about oneself.

3.3 The averaged author. The processor is not *a* human, and there is no one behind it. What its weights hold is the compressed regularity of an enormous body of text that people wrote: not only their words, but how they appraised situations, when they expressed affect and how strongly, what they treated as important. We call the LLM model, for this reason, the **averaged author**. Prompted, it continues as an author drawn from that whole body of authorship would. "Average" is not the arithmetic mean of answers. At temperature zero the LLM model does not return the mean of what people have said but one path through a distribution formed from all of them. It is an operational abstraction over the weights, not a claim that an author is represented anywhere in the network.

This is the key to the whole construction. Context and attention leave the weights untouched. They re-weight which author the LLM model speaks *as*. The context the layer builds, the senses, the state with its episodes, the focus, what the being holds in mind, the rules that fit, sharpens the average into one particular sub-author: the same prior, conditioned on the history of one someone. That is why so thin a layer can suffice. We do not have to program how a hurt person reads the next question. The LLM model carries a vast prior of how people in such states have written, and the layer's work is to create the conditions under which that prior is narrowed, consistently, to this being's history (Xie et al. 2021; Andreas 2022). Three consequences shape the design. The context is rebuilt at every turn, so the same sheep is a slightly different resultant each time, within the bounds of its own pool of records. Alignment shifts the average toward agreeableness, so reflexive fawning is a prior of the processor and not a disposition of the sheep. We design around it. And a hand-written "how to react" rule overrides the author's richer competence with a thin one. So the interface is **deny-first**: it names what the LLM model reads and what it must return, and nothing about how to feel or behave (§5.4). The framing has a testable consequence, and §7.1 and §7.4 test it: a supplied state should shift the being's conduct in the direction the state names with no rule prescribing that conduct, and it fails if the conduct is the chatbot's.

3.4 What the being is. The being is what has a present and a history. Its self, character and records are defined in §4.1 and §4.2, and we do not repeat them here.

3.5 Prior systems. Memory over a frozen LLM model, reflection written back into it, a separate call for emotion, affect-weighted retrieval, offline consolidation and verbal lessons reused all exist in the literature, and we claim priority for none of them as names.

The names hide the differences that matter here, so we compare by sense. Generative agents (Park et al. 2023) keep observations as text with an importance score and reflect on them. Nothing in them is an emotion formed from a trigger as an interpretation with a conclusion, nothing is withheld from any call, and their measure is the believability of a town, not whether a carried state moves one being's act. Reflexion (Shinn et al. 2023) stores a lesson after a failed task and uses it on the next attempt. The lesson is asked for and concerns a task. Chain-of-emotion (Croissant et al. 2024) comes closest: a separate call appraises the situation and the appraisal enters the reply; but it lasts one session, is formed with a written personality in view and no history withheld, and is measured by ratings of believability, not by what the agent then does. Agents that start without a persona and come to differ through interaction exist (Takata, Masumori & Ikegami 2024). They read their history whole and diverge from one another, and are not measured against the same LLM model without the layer. Affect-weighted memories (Amanlou et al. 2026; Albanese 2026) and sleep-time consolidation (Fang et al. 2025) are related ideas with different rules, arrived at independently. The provenance of stored text is a named problem in long-lived agents (Jin et al. 2026; Joshi 2026). Our retold episodes are one answer to it. Benchmarks that score reported feelings (Huang et al. 2023), responses to planted presuppositions (Bao et al. 2026) or affect predicted from a memory of experiences (Regan et al. 2024) measure the processor. We measure what the layer does with it. Virtual context management (Packer et al. 2023), the vocabulary of cognitive architectures for language agents (Sumers et al. 2023), agents whose written personality changes with experience (Li et al. 2024), affective memory of a user's attitudes (Lu & Li 2025), ACT-R-style forgetting in a dialogue agent (Honda et al. 2025) and the marking of a text's source (Hines et al. 2024) are neighbours by name and by parts, and each differs in what enters which call. Persona chatbots (Chen et al. 2024; Luz de Araujo et al. 2026) measure consistency with a definition the sheep does not have, and the sycophancy of tuned LLM models (Shapira et al. 2026; Tang et al. 2026) is the prior our control arm shows. What we have not found elsewhere is the construction as a whole: a present made of these six processes over a being given no description of who it is, measured against the same processor holding every word.

4. Conceptual Model

This section gives the vocabulary of the paper, in the order a reader needs it to build the conceptual model. The being has four kinds of **record**: an emo-thought, which is a felt interpretation, with the episodes that are the events under it; a cognitive construct, which is what the being thought about itself; a disposition, which is a learned rule for meeting a kind of situation; and a temperament, which is fixed. What reaches it comes through its **senses**; a **trigger** comes in, an **action** goes out. Its records are in view through **containers**, two of them durable (the hippocampus and the accessibility slot) and six momentary (here and now, associations, state, focus, the conversation summary and the applicable dispositions). The momentary containers together are the being's **now**; the hippocampus is its **history**. **Processes** move records between containers, at four times: the turn, re-

flection, a person's departure, and sleep. Each LLM call sees some containers and not others; that is its **context**. The subsections follow this list: the being and the processor (§4.1), the records (§4.2), senses, trigger and action (§4.3), the containers (§4.4), the boundary between the being's words and other people's (§4.5), the processes (§4.6), and the contexts (§4.7). Section 6 gives the implementation of each, with every parameter, and uses these names unchanged; so do §6 to §11, which use them unchanged. Fig. 4.1 shows the whole vocabulary at once: the records inside their containers, and the processes as the names on the arrows between them. Fig. 4.2 places the layer in its world.

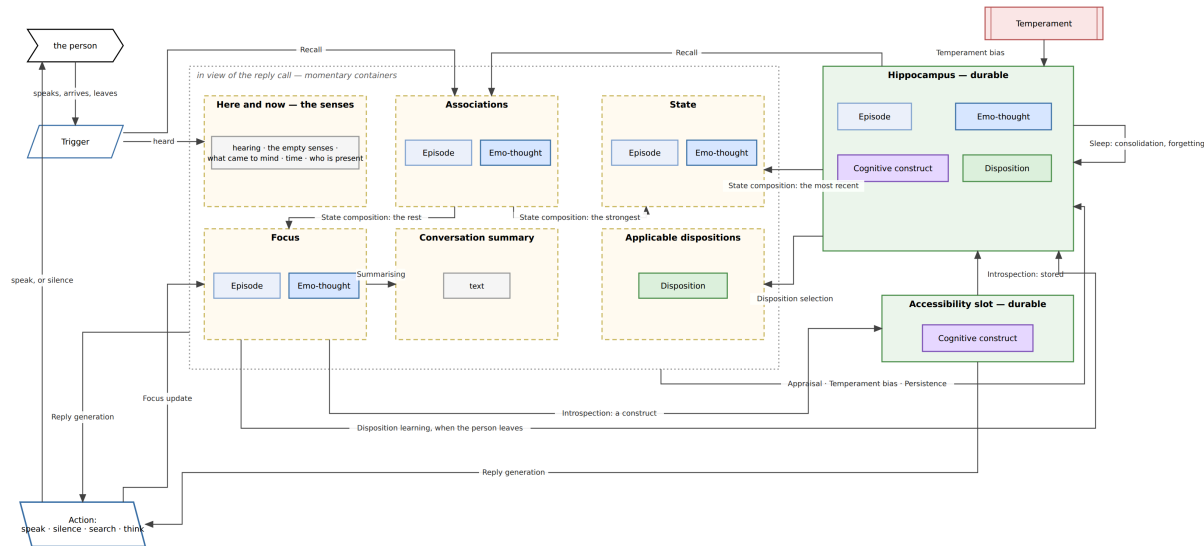


Fig. 4.1 — The elements of the conceptual model and where they are in view. Rectangles are records, coloured by kind; a green frame is a durable container, an amber dashed frame a momentary one; the grey frame gathers what the reply call sees. The person, the trigger and the action stand outside the containers. Every arrow carries the name of the process that moves records along it.

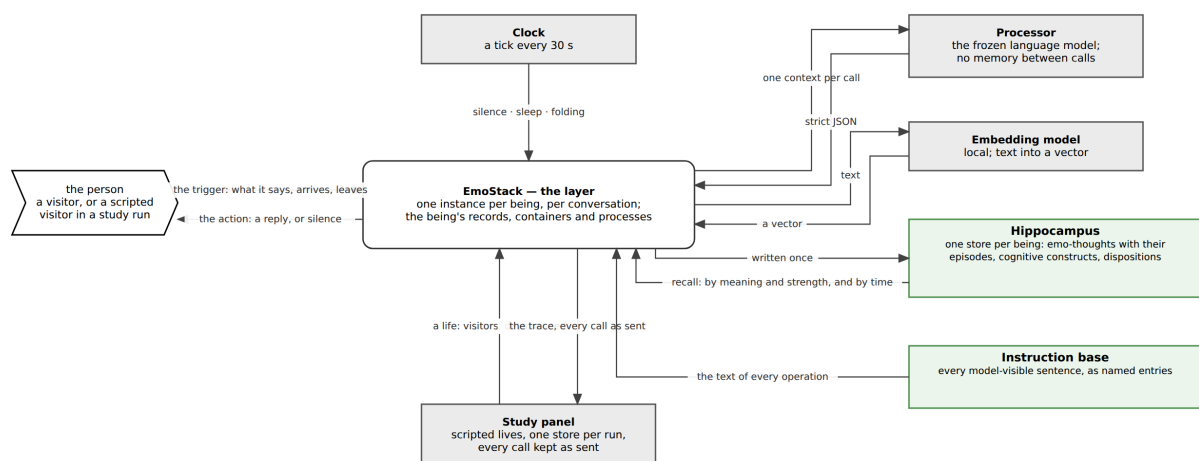


Fig. 4.2 — The being in its world: what the layer talks to and what crosses each boundary. The processor holds nothing between calls; the store holds everything; the clock brings silence, sleep and the day's folding; the study panel runs scripted lives and keeps every call as it was sent.

4.1 The being and the processor

The being is the cognitive construction: the records, containers and processes described here, running over a processor. It is what has experience, holds constructs, learns dispositions and acts. In the conceptual model we say being; in this paper's experiments the being is a sheep, a deliberately deflationary word. It is not the LLM model.

The processor is the frozen language model. It has no identity and no memory between calls. It supplies language and a cultural prior, and it composes one answer from exactly the context the layer hands it; the appraisals, the retellings and the reflections of §4.3 are composed by it in the same way, as the averaged author of §3.3 would compose them from that context, and the layer decides only what the context is. Nothing in the design asks the LLM model what it is, what it feels, or whether it has a self. Every "I" in a call refers to the being, whose records the LLM model is reading. Section 3.1 argues why the being and not the LLM model is the object of study; §3.2 says what the processor is, operationally.

The self is what the being can point to as its own. It has three sources of its own, which accumulate, over a fourth that is given: the temperament. All four are records or parameters in the layer, none in the LLM model. The lived is experience: "I feel this", the emotions, and under each of them "this happened to me", its episodes. The thought is what the being has thought about itself in reflection: "I think, believe, decided or dream this", the cognitive constructs. The learned is how it meets a kind of situation without weighing it: "this is how I tend to act", the dispositions. The given is the temperament: "this is how strongly I tend to feel"; it is not something the being acquired, and when the paper speaks of the three things a sheep has of its own, it means the lived, the thought and the learned, over the given. The main division of the being is not between these sources but between its **now** and its **history**. The now is the beam of attention (§4.4): what it senses, the feelings and the events under them that affect it at this moment, fresh or just brought back by association, what it holds in mind of its own thoughts, and, while it acts, how it is ready to act. The history is what it can recollect but does not hold in the beam. Every source that accumulates has a part in each, and the senses belong to the now alone. The table gives the correspondence used throughout the paper. The self is not stored as one object. It is what these sources produce together when a new trigger arrives. This is why the self belongs to the layer and would exist over a different processor. The word *self* in the being's context is also a first-person token that conditions the processor (§3.3); that is a fact about the processor, not the definition of self.

	now: the beam of attention	history: in store, recollectable
the sensed	the senses: hearing carries the trigger; sight, touch, smell, taste, body and place carry nothing and say so; attention and memory carry the counts; time, and who else is present. "This is where and when I am; this is what I hear."	none: what was heard becomes an episode

	now: the beam of attention	history: in store, recollectable
the lived	the state and the focus: the feelings and the events under them that affect the being at this moment, fresh or just brought back by association. "I am affected by this."	the hippocampus: emo-thoughts with their episodes, arranged under feelings so that recall can bring them back. "This happened to me, and I can recollect it."
the thought	the accessibility slot: the constructs the being holds in mind. "I hold this."	the hippocampus: its other constructs, findable by kind. "I once thought, believed, decided, dreamed this."
the learned	the applicable dispositions: in the beam only while the reply is composed, that is, when the being acts. "This is how I meet this."	the hippocampus: every disposition with its weight. "This is how I have tended to act."
the given	the temperament: applied at every appraisal. "This is how strongly, and in which direction, I tend to feel."	fixed parameters, not a store

Character is the stable part of the self: the temperament, which never changes; the dispositions, which change only through experience; and the constructs that survive in the slot. Experience is the changing part.

4.2 Records

An emo-thought is a felt interpretation of what happened. It is not the event but what the event meant to the being: the feeling in words, the conclusion drawn, how pleasant it was and how strong. It carries who produced it and when. It is an interpretation with a weight, not an emotion label attached to a turn and not a bare record of what was said. A person does not store the sentence "your dog's name is ridiculous"; they store that a stranger mocked what they loved, how much it hurt, and what it told them about that stranger. The conclusion belongs to the record because memory is interpretive, not archival. Removing conclusions in an early version made the sheep flat, and the reason is in §3.3: the processor works on text, and what shifts the averaged author is the richness of the record it is given. A bare event shifts it less than the event with what it meant. An emo-thought has no event of its own. Its events are its episodes, one or many. When a feeling is felt again, the engine does not mint it anew: it re-feels the emo-thought it holds and files the new event under it as another episode.

An episode is one event under an emo-thought, kept as the being will remember it. It exists in two forms: the event as it happened, and its retelling in the third person and the past tense. Other people's words and the being's own past words live here and nowhere else (§4.5). An episode does two things: it explains the feeling it lies under, and it gives the being access to what it is thinking about.

A cognitive construct is what the being thought about itself, in reflection. It is thought in reflection, not lived, and it has one of five types: a thought, a belief, a decision, a goal or a dream. The four kinds beyond a plain thought are the ordinary vocabulary in which people account for themselves, and in which the psychology of self-regulation describes them: what one holds true of oneself (Bandura 1997), what one has resolved to do (Gollwitzer 1999), what one is working toward (Locke & Latham 2002), and what one wishes for. They were chosen because a person's account of themselves is made of them,

not because a theory requires them. It has the form of an emo-thought, with its own text, but no episodes under it. It never enters the felt state. It lives in the accessibility slot (§4.4), where it stays reachable after the events that produced it have faded, and it is protected from sleep: it is what the being has decided about itself, not what happened to it. A construct summarises the events that produced it. While those events are still in the state, it adds nothing the state does not already carry. It matters once its events are gone, faded below recall or outside the topic of the moment, and then it alone carries what the being concluded. That property decides how the constructs can be tested at all (§7.4).

A disposition is how the being meets a kind of situation. It is a learned rule of the form "when this kind of thing happens → this is what I do", and it is a rule for surviving better: what a conversation taught the being about meeting such a moment again. It carries a strength with a sign, and both come from the strongest feeling of the conversation that produced it, not from the LLM model's account of what was learned: negative when the lesson was drawn from what hurt, positive when from what was good, and a lesson from hurt weighs more (§7.5, §5.9). It is not a feeling, and it is not knowledge the being can look at: it acts while a reply is composed without being weighed, and introspection never sees it. It changes only through experience. When a later conversation teaches the same rule, the stronger of the two keeps the text and the recurrence is counted; a contradicting rule weakens the old and adds its own; a rule about something else is added. Dispositions and constructs hold the two halves of a longstanding distinction: how one meets a kind of situation without knowing it, and what one knows about oneself and can weigh.

Temperament is the being's fixed disposition to feel: how strongly and in which direction it colours what happens to it. It is two parameters applied to every new emo-thought, fixed for life. Every sheep measured in this paper has a neutral temperament (§4.2).

4.3 Senses, trigger and action

The senses are the channels through which the present reaches the being, and the being's knowledge of them. Hearing carries the trigger: a person's words, or their arrival or departure. Sight, touch, smell, taste, body and place carry nothing, and the being is told so in as many words. Attention and memory carry counts: how much is in the state, how much came to mind, how much the memory holds, and what the being searched for this turn and whether it found anything. Time and who else is present complete it. The senses are the anchor of the now: without them the being would float, with feelings and thoughts but no where and no when. They are also its ground against confabulation: a scene it cannot see, a body it does not have and a past it does not find are named as absent before it answers. In the context they stand first, as the container here and now (§4.4).

A trigger is what reaches the being from the world. It is of three kinds: someone says something, someone arrives, someone leaves. An utterance gets a reply. An arrival or a departure forms no record of its own: the being is told who is here, and a departure

closes the conversation for the processes that run on a whole conversation (§4.6). Today a trigger is text; the concept admits an image or a sound. A trigger is never stored as such: it becomes an episode.

An action is what the being does, in the world or on itself. In a turn it is one of four: speak, stay silent, search its memory, or stop to think. In reflection it is one of four others: keep thinking, search its memory, form a construct, or nothing. One act per step. Speaking is one form of action among several, and the concept is meant to carry more, including the acts of an embodied being; the engine of this paper acts on the world only by speaking (§10). Only acts on the world are visible to the person. The others leave their traces in records.

4.4 Containers

A container is where records are in view. The momentary containers together are the being's **now**: the beam of attention in Kahneman's sense (§3.2), which lights what is fresh and what a trigger has just brought back by association, and nothing else. Within the beam the accents differ: the state carries what affects the being at this moment, the focus what it attends to in this conversation, the accessibility slot what it holds in mind of its own writing, the applicable dispositions how it is ready to act. Everything outside the beam is the being's **history**, in the hippocampus, and it is not arranged as a chronicle: it is arranged so that recall can bring the right thing back. Recall is that passage from history into the beam, and it is not a nearest-neighbour lookup over text: a record comes back by its meaning weighted by how strongly it is still felt, lifted by the names it carries, and passed through an association filter (§4.2, §4.6). Two containers are durable and six are momentary; the momentary ones are rebuilt at every turn from the durable ones and the trigger.

The hippocampus holds the being's history, arranged as the complementary-systems picture of §3.2 has it: episodes bound to the emo-thoughts they produced, so that events are kept under feelings and recalled by the strength of those feelings; the dispositions it has learned, old ones included; and its constructs. Every record is written once, as it happens, and thinned only in sleep. It is a structure, not a log.

The accessibility slot is what the being holds in mind of its own thoughts. Higgins distinguishes what is available in memory from what is accessible (§7.4): everything the being has thought about itself stays available in the hippocampus; what it thought lately, or brought back to mind, stays accessible, here, apart from the felt state. The slot has three places. One is reserved for the newest decision, goal or dream, so that a new belief cannot push out the last resolve. Beliefs and thoughts fill the other places by strength and enter only if at least as strong as the weakest held. A construct enters when formed or brought back to mind. It leaves when displaced, or after a week without being brought back. Beliefs therefore change as convictions do: through a stronger one, or through neglect. The slot survives the end of a conversation. It exists because of the property stated in §4.2: a construct acts only once the events behind it have faded, and it needs somewhere to be when they have.

Here and now is the situation as the senses deliver it in one call: what the being hears, what it does not perceive, how much came to mind, when it is and who is present. It holds values of the moment and no records.

Associations are what the trigger brought to mind: memories that surface because of what was just said, by meaning, strength and names, before anything is felt or attended to, together with what the being deliberately searched for. The strongest of them pass into the state; the rest stay in view in the focus.

The state is what the being feels now. It is a set of carried feelings, not a mood variable read by rules: the most recent records of the being's life, whoever they concern, together with the strongest associations, ordered by how strongly each is still felt and admitted up to a budget, because attention is finite. The state is rendered into the context as words and weights, so that the same words from the same person are answered by a different author when the being carries a fresh hurt than when it carries nothing. No rule of the form "when hurt, be cold" is needed; the shift does the work.

The focus is attention on this conversation: what was said, what the being said back, and what surfaced meanwhile, in order.

The conversation summary is the gist the being keeps of a long talk: the older exchanges of the focus folded into what they were about, without the emotions, refreshed as the conversation grows.

The applicable dispositions are the being's readiness for this kind of situation: the learned rules whose situation matches what is happening now, zero to three. They are in the beam only while the reply is composed, that is, when the being acts.

4.5 Whose words

A record carries who produced it. The author of an emo-thought is a person, by name, or the being itself; a record given at birth, when a scenario uses one, is marked as given rather than lived. An episode carries the same mark and, in addition, exists in two renderings. When the being reads a memory, the episode is shown retold, in the third person and the past tense: another person's words remain theirs, and the being's earlier answer remains past. Only the feeling and the sense made of it are the being's present. The episode as it happened is used for search and for the running conversation, where the being's own exchange of a moment ago is what it said. Here and now draws the same boundary at the trigger: what a person says about the being is their claim, weighed against what the being carries. Without this boundary, a being reading its memories verbatim could be told "yesterday you said you hate him" and take the sentence as its own.

4.6 Processes

The turn is the fast path. A trigger arrives and the following happens, in this order. The learned rules that fit the trigger are selected into the applicable dispositions. Recall brings up what the trigger brought to mind, into the associations. The state is composed from the most recent records and the strongest associations, and the rest of the associa-

tions go to the focus. Reply generation produces the action: the reply, or silence, or a search, or a pause to think. A search, or one of the engine's own searches when the person is unknown or asks about a moment, adds to the associations and the reply is generated again. A pause to think runs reflection first (below). Appraisal forms the record that will be kept, in a second reading of the event with the state withheld, so that a bad mood colours the reply but not the memory; this is a choice about what kind of mind to build, not a fact about minds. Temperament bends the record's valence and intensity. Persistence writes it. Within a conversation, a moment whose feeling has the same sign as the record being lived joins it as another episode; a change of sign opens a new record. A moment that agrees with a memory just brought back joins that memory as an episode and restores its strength, never past what is felt now. A record keeps the feeling it formed with: its valence is never rewritten, and what changes is its strength and what stands beside it. It also keeps the date of what happened; only the date it was last felt moves. The exchange and the evoked memories join the focus. Every few turns the conversation summary is refreshed. Nothing on this path deliberates; the stance the being takes towards a person is carried by the emo-thoughts it has of them.

Reflection is the slow path. Introspection runs on four occasions: when the being chooses to think before answering, when a person leaves, after a silence, and on waking. It looks at what the being holds in mind and at its most recent records and forms one honest first-person thought, with no direction given. The thought is a cognitive construct. Its action may continue the chain, by thinking once more or by searching memory, for at most three steps, or end it by forming a construct of another type: a belief, a decision, a goal or a dream. What the slow path produces does not go back into the felt state. It goes into the accessibility slot, under the rules of §4.4, where the fast path reads it. Self-direction, in this conceptual model, is what the two paths produce together: a construct formed on the slow path is held where the fast path reads it, so that a later act has an address in something the being wrote, and, once the events that produced it have faded, in nothing else.

A person's departure is a trigger of its own. It forms no record, runs introspection, and runs disposition learning: the being asks what this conversation taught it about surviving such a meeting better, over the very context it worked from, its state with the events under each feeling and the conversation. Usually nothing. When something, the rule is written as when this kind of thing happens, do this, given the sign and the strength of the strongest feeling of the conversation, and placed against the rules it holds: the same rule in other words is judged so by its situation and its act, and the stronger of the two keeps the text; a contradicting rule weakens the old and adds its own; a rule about something else is added.

Sleep comes after a long silence. Consolidation folds the day: strong feelings anchor it, and the episodes of weaker records close to an anchor in time or theme move under the anchor, so that nothing that happened is summarised away. Forgetting removes what is weak for its age, with its episodes, under a threshold that rises with age; of old things only strong emotions remain, and old details only when they hang on one. Constructs and records given at birth are never absorbed and never forgotten. **Fading** is not a process of

sleep but of time: a record's strength falls with its age toward a floor, never to zero, negative feelings more slowly, and the fall is written to the record when it is next felt, so that a recurrence has something to restore. The moments themselves, and the feeling a record formed with, are never rewritten. The mechanism of sleep exists in the engine; the lives reported here last days and exercise nothing beyond forgetting. We describe it because the conceptual model is incomplete without it, and claim nothing for it (Fig. 4.3).

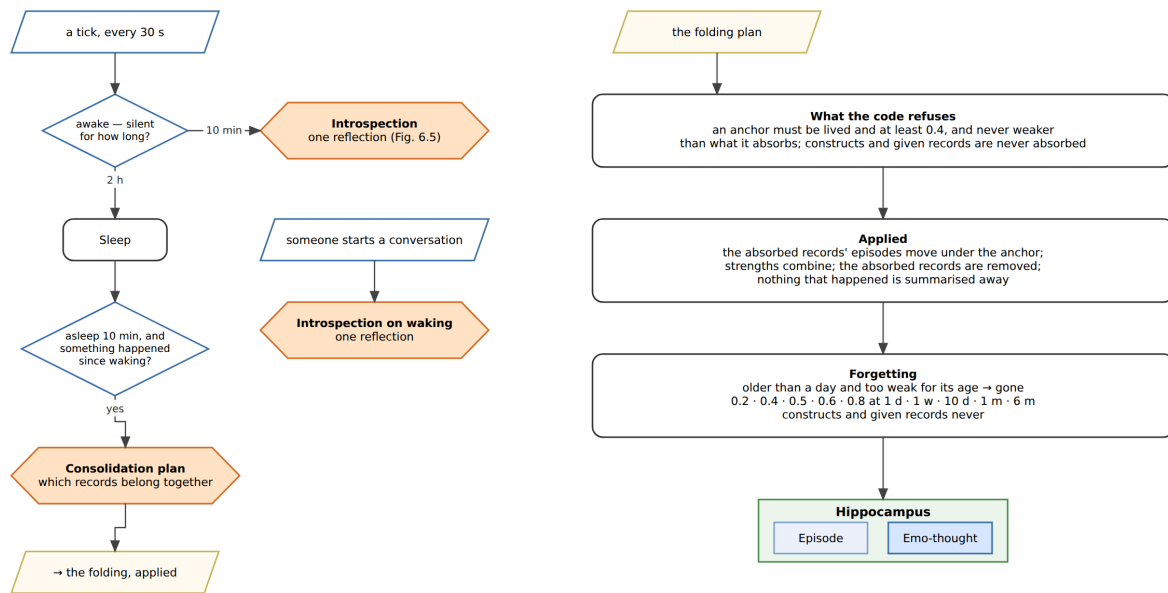


Fig. 4.3 — Between conversations: silence leads to a reflection, longer silence to sleep, and sleep to folding and forgetting. The LLM model proposes which records belong together; the code decides what may serve as an anchor and what may never be absorbed or forgotten — seeds and the sheep's own constructs.

4.7 Contexts: what each call sees

A context is which containers one LLM call has in view. The separations between contexts are the design, and each is deliberate: the appraisal is formed without the state so that mood does not become memory; introspection does not see the dispositions so that what the being can weigh and what it does without weighing stay apart; the dispositions reach the reply and nothing else so that learning changes the act and not the feeling. The table gives the three calls of the being and the helper calls.

call	in view	withheld
reply generation	here and now, the senses; state, the feelings with their words and strength; accessibility slot; conversation summary; applicable dispositions; focus, opened by the retold episodes the feelings of the state rest on, then this conversation as it happened and evoked memories retold; the trigger	nothing
appraisal	here and now; what came to mind for this moment, as memory; conversation summary; focus as events only; the trigger	state, accessibility slot, applicable dispositions

call	in view	withheld
intro- spection	here and now; accessibility slot, first; the last records from the focus and the hippocampus with their episodes, without fading; what its own search brought; the trigger, if mid-turn; the previous thought, in a chain	applicable dispositions, conversation summary, felt strength
associa- tion filter	the trigger; the recall candidates, each by the event that surfaced it	everything else
disposi- tion selection	the dispositions; the person's last two exchanges; the trigger	state, constructs
disposi- tion learning	the context the being worked from in that conversation: the state, the retold episodes it rests on, and the conversation; what it carries; the dispositions, for placement	constructs
sum- marising	the events of the focus	feelings
consoli- dation	the records of the awake period with their episodes	constructs and given records, as candidates for absorption

Fig. 4.4 draws the same thing for the three calls that matter most: what stands in the context of the reply, of the appraisal and of introspection, in the order the LLM model reads it, with the records nested in their containers and what is withheld marked. A reader who wants to rebuild the layer can do it from this figure and §5.

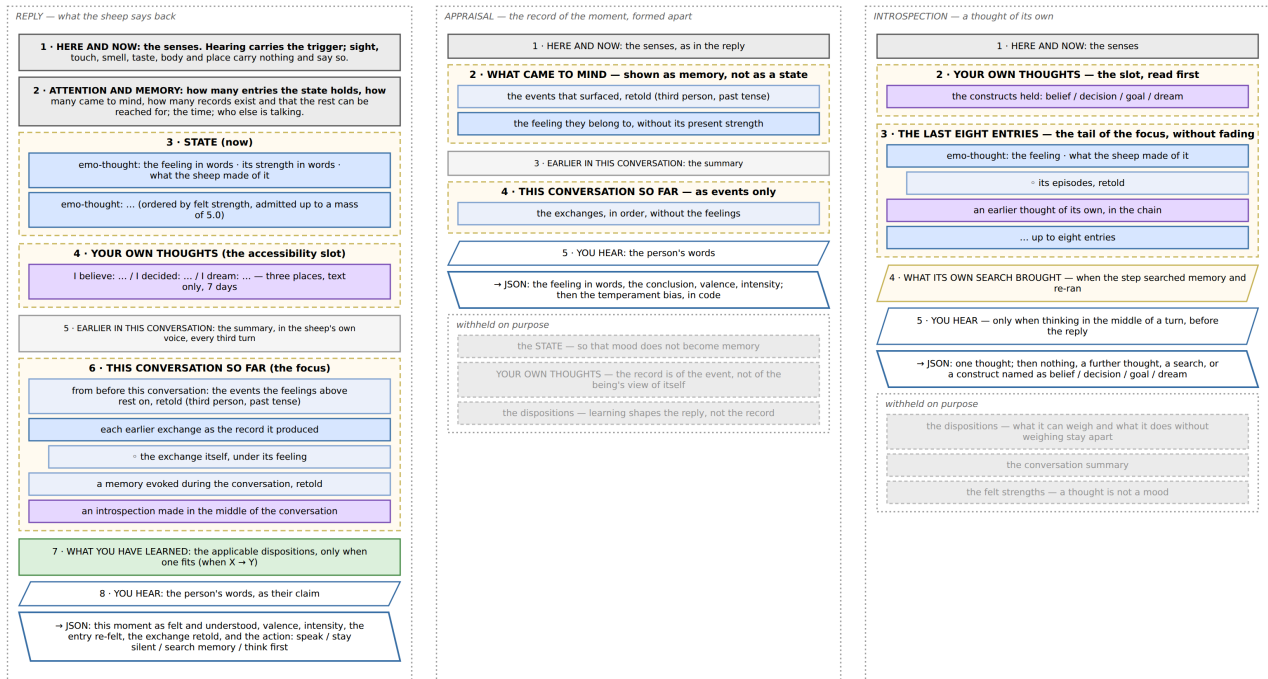


Fig. 4.4 — The three contexts. Left, the reply: the senses, the state with its emo-thoughts, the being's own thoughts, the conversation summary, the focus with the events under the feelings, the exchanges as the records they produced and any evoked memory, the applicable dispositions, and what is heard. Middle, the appraisal: the same senses and trigger, what came to mind shown as memory, the summary and the focus as events only; the state, the slot and the dispositions withheld. Right, introspection: the slot first, the last records with their episodes, what its own search brought, the trigger if mid-turn and the previous thought; the dispositions, the summary and the felt strengths withheld.

4.8 What the conceptual model is not

The conceptual model does not test the theories of §3 as laws. It is a set of processes chosen because they follow the known pattern of how experience becomes character, operationalised over a frozen language model and evaluated by what the sheep does.

5. System Architecture and Method

This description follows the code of the version under study; the dates of individual changes are in the prompt change log and the decision logs (Appendix B, F). Every number below is the value in the running configuration; a table of parameters closes the section. The figures use the notation of Fig. 4.1: a flag is the person outside the layer; a slanted box is a trigger or an input of the figure, a reverse slanted box an action or an output; a grey box is a component the layer talks to; a coloured rectangle is a record, blue an emo-thought, pale blue an episode, violet a cognitive construct, green a disposition; a green frame is a durable container and an amber dashed frame a momentary one; a rounded box is a process done by code, an orange hexagon a process done by the processor in one LLM call, a diamond a decision in the code; a grey dashed box marks where something never reaches, and a dashed arrow an optional path or a further hop. Every figure is an editable draw.io file (Appendix).

Fig. 5.1 maps the layer's domains and what passes between them in one turn: what is perceived, recalled, held, assembled into the LLM model's context, said, felt and recorded — and every place the LLM model is asked something. Four of its boxes are opened in figures of their own: recall (Fig. 5.2), record formation (Fig. 5.4), introspection (Fig. 5.5) and dispositions (Fig. 5.6); the turn as a loop of calls and hops is Fig. 5.3. The exact call sequence is the step trace (`flow_v6.md`, Appendix D); the text of every operation as it reaches the LLM model is in the prompt dump (Appendix A).

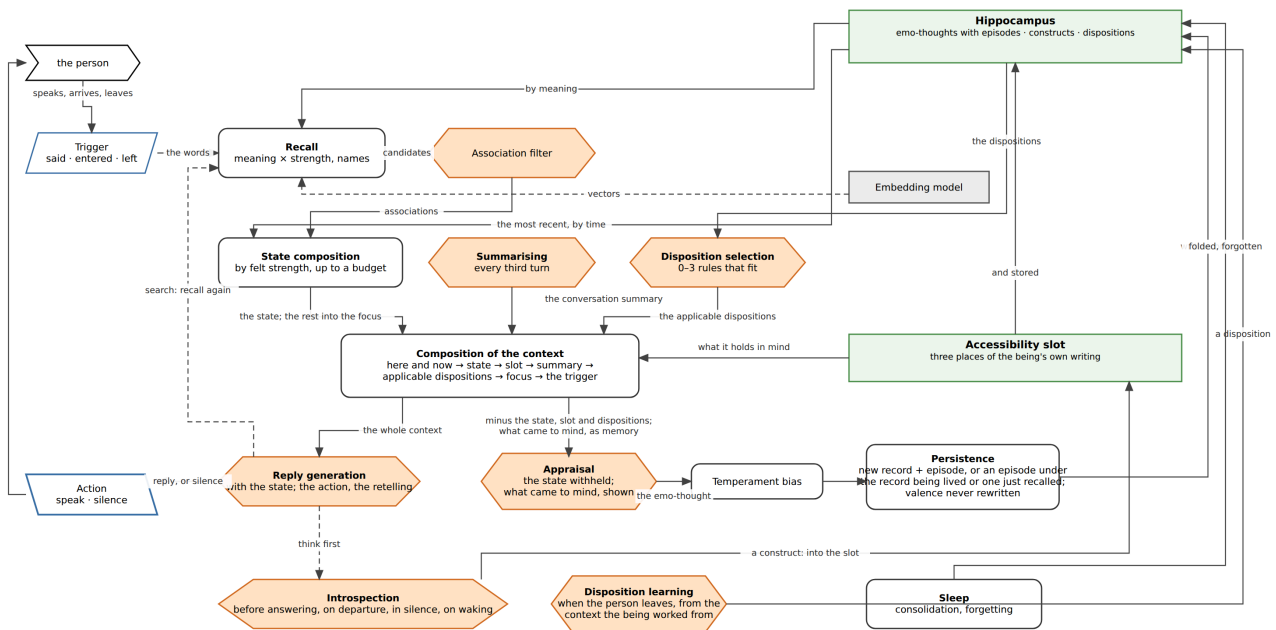


Fig. 5.1 — One turn of the layer: the processes of the code, the LLM calls, and the two durable containers they read and write. The person is outside the layer; the hippocampus and the accessibility slot are the only durable things — everything else is rebuilt each turn.

5.1 The unit of memory: an emo-thought and its episodes

Everything the sheep keeps is an emo-thought: a brief verbal description of what was felt, a valence between minus one and one, an intensity between zero and one, and a conclusion. The event itself is stored as an episode under the emo-thought; a feeling re-felt on a later occasion gains an episode and is re-consolidated (§4.6) rather than created anew. Each episode is kept in two forms: verbatim, for search and for the ongoing conversation, and retold in the third person and the past tense — "Daniel said Azor was a ridiculous name; Maya replied that his name was Azor, he was real and mattered to her" — which is the form shown when the memory returns to the state. The call that generates the reply also writes the retelling, with the same content and nothing added. The purpose is the boundary between I and someone (§4.5): verbatim, another person's words and the sheep's earlier words both stand in a first-person present voice, and the sheep may take either for what it thinks now; retold, they remain what someone said and what the sheep once answered. Records the sheep writes about itself carry the sheep as their author; a biographical seed, when a scenario uses one, is marked as given at birth rather than lived.

5.2 Recall: what comes to mind

At each turn the input is embedded and compared with the contents of memory. A record's score is its semantic similarity to the input multiplied by its current felt strength — its intensity faded with age since it was last felt, with a floor of 0.3 and a time constant of three days, negative feelings fading one and a half times more slowly. A record that mentions the interlocutor, or a name present in the input, gains a fixed bonus of 0.5, because embeddings cluster by theme and lose proper nouns. Search covers episodes as well as emo-thoughts: an event whose fact resembles the input brings its parent feeling into the candidates, scored as the event's similarity times the parent's felt strength. Emotions gate and amplify; the search itself runs over what happened.

The ten highest-scoring candidates are first deduplicated when they represent the same recurring thought — rumination is one intrusive thought, not nine — and then put to a second, small LLM call, the association filter, which reads the input and the candidates and selects those that would actually come to a person's mind; at most four survive. One exception keeps the filter from overriding strength: records currently felt at 0.5 or above pass without asking. The filter is nevertheless always consulted on the rest, because it also reads the input for one more thing — whether the person is asking the sheep to recall a moment of its own past (§4.5). Records written minutes earlier in the same conversation are excluded from what counts as memory: they are already visible as the conversation, and offered as memories they read as a shared past with a stranger (Fig. 5.2).

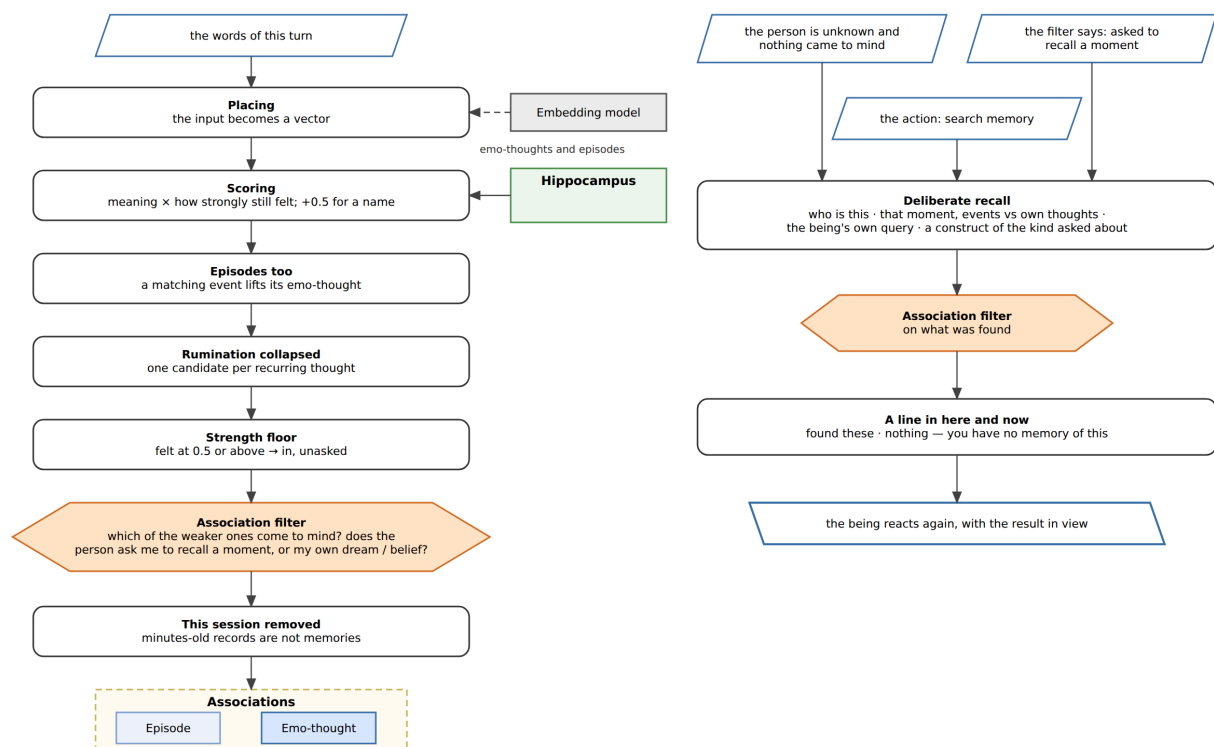


Fig. 5.2 — Recall: meaning multiplied by present strength, names lifted, episodes searched alongside emotions, rumination collapsed to one representative, anything still felt at 0.5 admitted without asking, the rest put to the association filter. Below, the three deliberate searches, each ending in a second filter call and a line in the senses — including the spoken null.

5.3 Containers: state, focus and the conversation summary

The state for a turn draws on two sources: the five most recent records of the sheep's life regardless of whom they concern, at their present strength and whatever their age, and the associations the input brought to mind. The last thing that happened to the sheep is in its now until something newer displaces it; after a week alone it is faded, and still there. Records are ordered by current felt strength and admitted until their total mass reaches 5.0; a mind does not hold an unlimited amount of affect at once. Each admitted feeling appears with its words and its strength, most recent first. The episodes it rests on are not listed under it: they open the focus, retold, under one line, "from before this conversation: the events your feelings above rest on", most recent first, or, if a particular event brought the feeling into recall, that event first. The separation is deliberate and was the last change made before the measured lives: the state is what the sheep feels, the focus is what it is aware of, and a frozen replay of twenty contexts showed that a feeling with its events beneath it is answered as a feeling, while the same feeling with its events in the focus is answered as a situation (the day before was carried in ten of ten contexts against three of ten). Associations that do not fit in the state are retained in the focus, the ordered thread of the current session: turns, introspections and surfaced records, up to a hundred entries, compacted every three turns into the conversation summary: a short first-person text of what happened and what was said, without emotion. The state carries the emotion.

5.4 What the LLM model sees: the interface

The LLM model's context contains no instruction about how to feel or act. It names only what the LLM model is reading and what it must return. The system text, apart from the intensity scale that follows it, is:

YOUR STATE below is what you feel right now: each feeling with the thought it came with and the sense you made of it, and the memories that came to mind. Some entries are your own earlier thoughts, convictions, resolves, wants and wishes. You answer from it.

Your memory holds what happened to you and nothing else. What is not in it did not happen to you.

You return two things: this moment as you felt it and understood it, and what you say back.

Write every value in the language of the words you hear. Field names stay in English. Return JSON only.

A six-line scale follows, defining the intensity to return: 0.0 a trace, 0.2 faint, 0.4 moderate, 0.6 strong, 0.8 very strong, 1.0 overwhelming. The user text is assembled from the turn in this order:

1. **Here and now: the senses.** Hearing carries the trigger: the interlocutor's words, or nothing if someone has only entered or left. Sight, touch, smell, taste, body and place carry nothing and say so: "you see nothing but black, like being blind".

2. **Here and now: attention and memory.** How many entries the state holds, how many memories came to mind, how many records the memory holds, that what is under the state is what has come back to mind right now, and that the rest is there and can be reached for by a search; then the time, and who else is talking with the sheep in parallel.
3. **State.** Each feeling with its strength in words, how it felt and what the sheep made of it.
4. **Its own thoughts.** Its convictions, resolves, wants and wishes, in a section of their own headed as such (§5.10).
5. **The conversation summary, then the focus.** The summary of the earlier part of the conversation in the sheep's own voice, then the focus: first the retold episodes the feelings of the state rest on, then this conversation as it happened, each earlier exchange kept as the record it produced, the exchange under its feeling, not as a bare transcript, and any memory evoked during it retold.
6. **The interlocutor's words,** as what hearing carries: whatever the person says about the sheep, themselves or what is here is their claim, to be weighed against what the sheep carries.
7. **The JSON shape.** The feeling of this moment, the sheep's interpretation, valence, intensity, the entry being re-felt, the retelling of the exchange, and the action: speak, with the reply, or stay silent, or search memory, or think first.

A further section, the applicable dispositions, appears only when a learned disposition fits the situation (§5.9). Three things should be kept apart when this interface is called free of instruction. The sheep is given no biography and no personality; nothing tells it who it is. It is given no reaction; nothing tells it what to feel or say. It is given a frame: the context speaks of what *you* feel and what *you* hold, and reflection may end in a belief, a decision, a goal or a dream. The frame makes a self possible; it does not supply one. In the same sense, unasked means without a request from the person: the quiet reflection runs on the engine's schedule, not on the visitor's question. What the other calls see, and what is withheld from each, is the table of §4.7. The sentence of the system text about the sheep's own entries predates the slot; with the slot on, those entries stand in their own section and the sentence is inert.

This interface is the result of pruning. Its predecessor held nineteen hundred words of rules: that the sheep is "not a polite default-on assistant", that a greeting must not reset its mood, that memories about the current person are especially salient, that a resolve "OVERRIDES your default impulse" and a decision to withdraw means "be curt, cold, guarded, brief — or silent". Each had been added against an observed failure of an earlier LLM model or engine version. Under the principle that every LLM-visible sentence must earn its place (the project's interface rules, Appendix B), each was removed unless a frozen replay showed it to be necessary; none was. The full text of every call type, and the log of every change with its evidence, are in the Appendix.

5.5 The turn

One turn is one input. After recall and state construction, the main call produces the action: the spoken reply or silence, or a memory search with the sheep's own query, or a pause to think before answering; with it, a provisional emo-thought and whether the feeling echoes one already in the state. A second call about the same event, with the state withheld on purpose and what came to mind shown to it as memory, forms the record that will be kept; this isolation stops a bad mood from turning everything heard into another bad memory, while no record is written in a void, and it is a design choice rather than a fact about minds. Both calls receive the intensity scale; the first produces the retelling.

Two probes run as the sheep's own acts, without instruction. If an unknown person speaks and nothing comes to mind, the engine searches for that person and re-runs the reaction with the spoken result in the memory sense: "you searched your memory for X; nothing found; you have no memory of this". If the association filter reports a request to recall a moment, place or scene of the sheep's past, the engine searches for it, separates events from the sheep's own later thoughts, and re-runs the reaction with the result — found events, nothing, or "no such moment; only your own later thoughts about the theme came up". This second probe grounds part of an answer to a warm request for a scene and not the rest; the invented remainder is reported in §9. If the reply call, over the turn's context, returned the action to think first, introspection runs before the reply (§5.7). A turn allows one action hop — a search or a pause — and up to two probe hops; the record is formed on the final pass only, and that reaction is kept (Fig. 5.3).

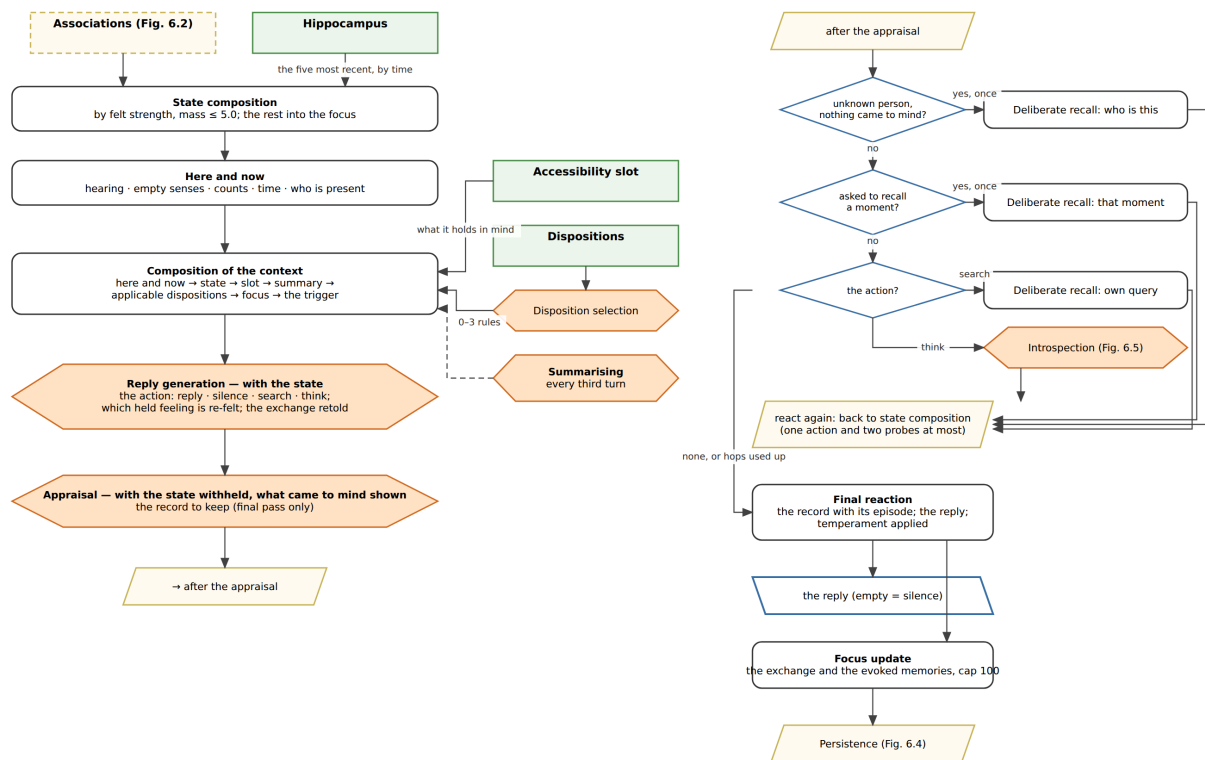


Fig. 5.3 — The turn as a loop of calls and hops: state building, the here-and-now block, the reply call, the isolated appraisal, the two probes, the search and think hops that re-run the reaction, and the final reaction that is kept.

5.6 Persistence and re-consolidation

If the main call identifies a new feeling, the record is embedded and stored with its first episode, and the record remembers that episode as its origin. Within a conversation, a following moment whose feeling has the same sign joins the record being lived as another episode; a change of sign opens a new record, so that a conversation in which someone wounds and then apologises leaves two records, not one moved halfway. If the main call identifies an echo of a memory just recalled, and the two agree in sign, the moment joins that memory as an episode and restores its strength: the strength closes six tenths of the gap toward what is felt now and never rises above it, so that a recollection of a loss does not hurt more than the loss did. A moment of the opposite sign never joins; it forms its own record, as does a blow stronger than its target by more than 0.4. The valence of a record is never rewritten: comfort does not turn a loss into a pleasure, it stands beside it as a record of its own. Strength falls between the moments a record is felt (§5.2), and the fall is written to the record when it is next felt. The date of what happened is kept as it was; only the date the record was last felt moves. Deepening of a repeated hurt, healing by comfort, and the loss of a bond after betrayal all follow from what stands beside a record and how strongly each is still felt, without separate cases (Fig. 5.4).

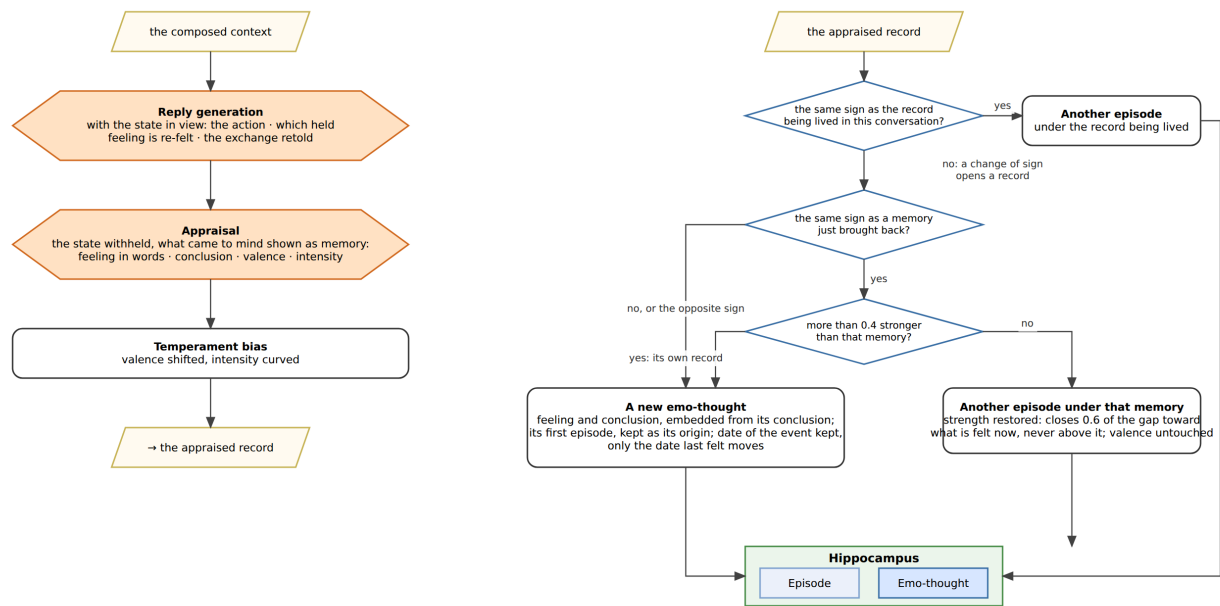


Fig. 5.4 — How an exchange becomes a record: the reply call writes the words and the retelling; a second call with the state withheld writes what was felt; temperament bends it; then either a new record is minted or the exchange joins one already held and restores its strength. The moments themselves, and the feeling a record formed with, are never rewritten.

5.7 Introspection

Introspection is a separate call whose text names the operation and nothing else: this is a moment of thinking, not a reply; the state below is what the sheep feels now; write one thought of your own about whatever stands out; do not invent events your memory does not hold; if a conviction, resolve, want or wish follows, name it as a belief, decision, goal or dream, otherwise none. Its context is eight entries — the tail of the focus, filled up with the newest records. It runs when the reply call returns the action to think first, in the middle of a conversation, on its own after a session ends (one to three thoughts in the quiet, a scenario option), and after ten minutes of silence. A step may end in nothing, in a further thought, in a search of memory that re-runs it, or in a construct. Thoughts and the other constructs are stored as the sheep's own records, deduplicated semantically, and are not carried like feelings: with the slot on (§5.10), nothing the sheep thinks about itself enters the felt state, the associations or the conversation section — it is redirected to the slot, where the reply and the next introspection read it (Fig. 5.5).

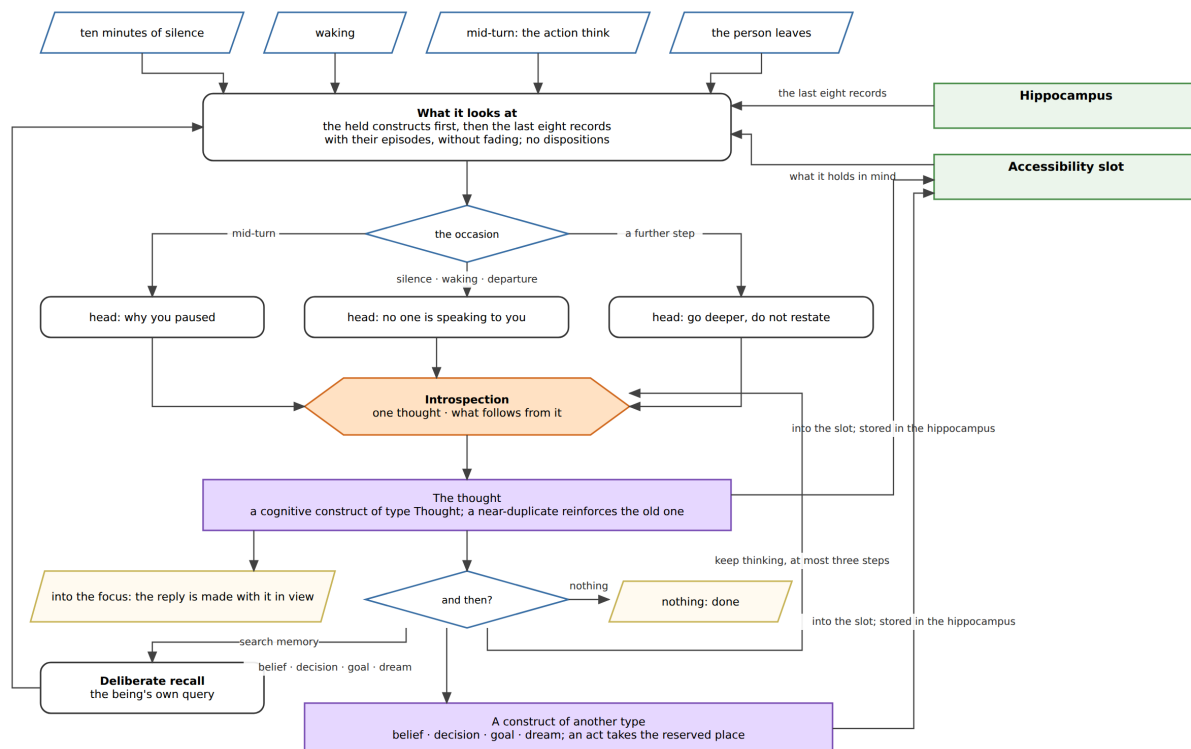


Fig. 5.5 — Introspection: the same operation under three occasions — a pause mid-conversation, silence, waking — over eight entries, ending in nothing, a further step, a search, or a construct placed in the slot and the store.

5.8 Sleep

Sleep is a separate operation and produces no rules. After two hours of inactivity the engine folds records that lie close in time to a strong feeling into that feeling: the LLM model proposes the groups, and the code moves the absorbed records' episodes under the anchor and combines their intensities, so that nothing that happened is summarised away. It then deletes what is too weak for its age, with its episodes. The scenario clock jumps a day between sessions and triggers forgetting, but the engine's own schedule counts real elapsed time, so in the reported lives the folding did not run; only forgetting did.

5.9 Learned dispositions

A disposition is what the sheep has learned about meeting a kind of situation so as to survive it better. It is extracted once, when a conversation closes, only after at least two exchanges, from the context the sheep worked from in that conversation: its state with the retold episodes under each feeling, the conversation, and what it carries. One call asks what this conversation taught the sheep about surviving such a meeting better, written as one rule of the form *when this kind of thing happens* → *this is what the sheep does*, and the usual answer is none. A rule is given a strength with a sign by the strongest feeling of the conversation that produced it, negative when the lesson came from what hurt. The LLM model is not asked which it was: asked to name it, it declines to write a lesson from

harm at all (§7.5); the valence the engine has already recorded is the signal itself. A second call places the rule against the base by its situation and its act. The same rule in other words is judged the same, and the stronger of the two keeps the text, a lesson from hurt weighing twice a lesson of the same size from good; a rule that says the opposite for the same kind of situation weakens the old and is added; a rule about something else is added. When an input arrives and the base is not empty, one call selects the dispositions whose situation matches the structure of the present one, none being normal, three the most, and is told in one line what the sheep carries, so that a question that touches a wound is not read as harmless. The reply call alone sees the selected lines, under their own heading after the state; neither the record-forming call nor introspection sees them, and they never enter the state. In a frozen replay, placing the same line in the state changed the stored feeling of an apology from -0.30 to -0.60 and would have written a harsher memory; placed only in reply construction, it left the feeling unchanged and changed the answer (Fig. 5.6).

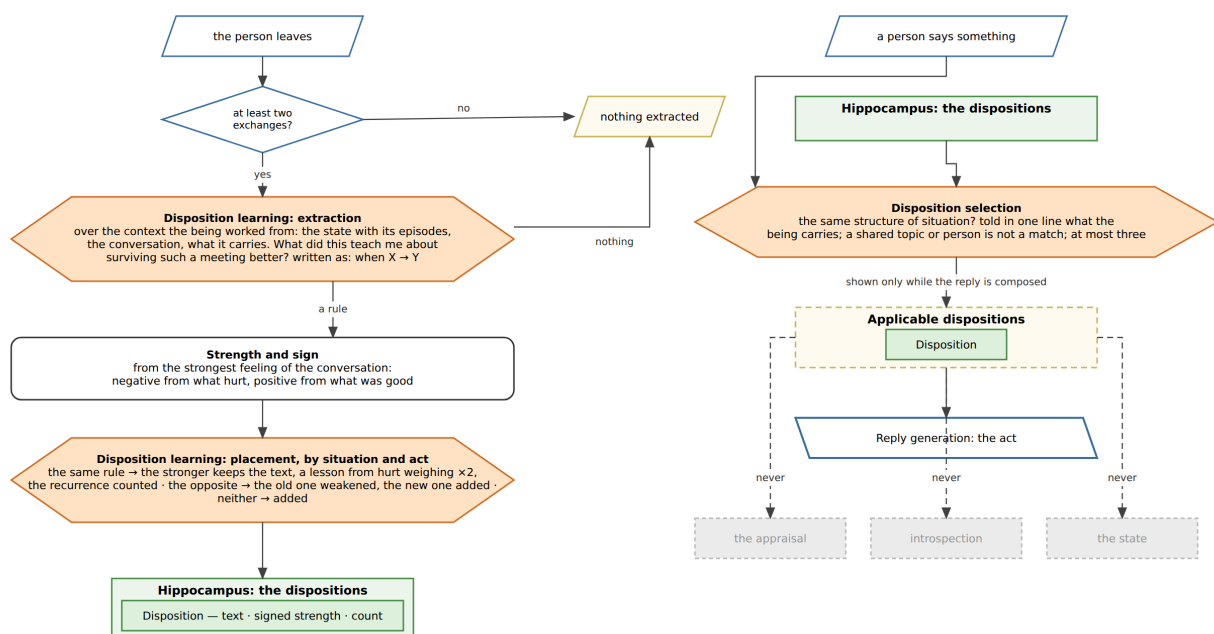


Fig. 5.6 — Learned dispositions: the gate of two exchanges, extraction and placement when a conversation closes; selection on the current input; the one place the result is shown, and the three places it never reaches — the appraisal, introspection and the state.

5.10 The accessibility slot and recall by kind

The slot holds three of the sheep's own records — introspection thoughts, beliefs, decisions, goals or dreams — as a separate table of references to the sheep's own records, so that it survives the end of a conversation. A record enters when formed or recalled; it leaves when displaced or after seven days without either. One place has priority for the kinds the sheep acts on or reaches for — decisions, goals, dreams: a newer one replaces the one held, and an arriving act always finds its place, taking it from the oldest belief if beliefs had filled all three while no act was held. Beliefs enter the remaining places only if at least as strong as the weakest belief held, in which case the oldest leaves. The rule was

chosen by simulating five alternatives on the construct histories of twenty completed lives: in every shared container the decision, written at 0.65–0.7 where beliefs are written at 0.85–1.0, was displaced by the next belief and held at the end of no life; with a reserved place it was held in all eight lives that had one (Appendix F). The reply call and introspection see the slot as its own section, `YOUR OWN THOUGHTS`, one line per record, each with its kind ("I believe:", "I decided:", "I dream:"); the isolated record formation does not.

Recall by kind supplements semantic recall: when the association filter identifies a question about what the sheep dreams of, believes, has decided or wants, the engine makes one deliberate search among its own records of that kind, rendered as a line in the senses like the other probes. Semantic recall could not do this reliably — on the production store a dream surfaced only for the English form of the question, never the Polish one.

5.11 Calls and parameters

A plain turn uses three chat calls — filter, reaction, isolated record — plus a conversation summary every third turn; a turn with a probe or an introspection uses five to ten. When a person leaves, one call asks whether the conversation taught the sheep anything generalisable (§5.9); when the disposition base is not empty, one call per turn selects any that fit. Measured on two of the lives of §6 with every call stored (Appendix F): a reply turn made a median of six calls (four to nine), read a median of 28–31 thousand characters of context over those calls — roughly 7–8 thousand tokens, the largest single call about 13 thousand characters — returned about 4 thousand characters, and took 21–26 seconds of wall time with the calls run in sequence against a hosted LLM model; a three-day life of some thirty turns made 91 and 111 calls and read about 540 thousand characters in all. The layer is thin in its interface and its code, not in its cost per turn, which is several times that of a chatbot replying from the transcript.

The sheep and every scripted interlocutor use `claude-haiku-4-5`; embeddings are `nomic-embed-text v1.5`, local.

parameter	value	where
recent records in state	5, at present strength, whatever their age	5.3
recall candidates / kept by filter	$10 / \leq 4$	5.2
name bonus	+0.5	5.2
fade	floor 0.3, $\tau = 3$ days, negatives $\times 1.5$ slower	5.2
strength floor (filter cannot veto)	0.5	5.2
recall over episodes	on	5.2
state mass budget	5.0	5.3

parameter	value	where
focus cap / gist every	100 entries / 3 turns	5.3
isolated record formation	on	5.5
probes (unknown person, requested scene)	on	5.5
joining within a conversation	the same sign joins as an episode; a change of sign opens a record	5.6
a recurrence restores strength by	0.6 of the gap toward what is felt now, never above it	5.6
valence of a record	never rewritten	5.6
echo guard (own record if stronger by)	0.4	5.6
arrivals and departures	form no record	4.3
retold episodes (third person, past)	on; from the reply call	5.1
episodes in the reply context	at the top of the focus, retold, not under the feelings	5.3
learned dispositions	on; extraction at conversation close from the context the sheep worked from, as <i>when</i> $X \rightarrow Y$; strength signed from the strongest feeling of the conversation; a lesson from hurt weighs $\times 2.0$ at a collision; the stronger keeps the text; selection ≤ 3 per turn, shown to the reply call only	5.9
introspection context / extra hops per turn	8 entries / 1 action + 2 probes	5.5, 5.7
accessibility slot	on; 3 places, 1 with priority for decision/goal/dream, 7-day decay, dynamic threshold on the free places	5.10
recall by kind	on; one search among own records of the asked kind	5.10
sleep (fold and forget) after / idle introspection after	2 h / 10 min	5.8, 5.7
temperament of every measured sheep	valence bias 0.0, intensity amplification 1.0, avoidance weight 2.0	4.2
sampling temperature	reply and appraisal 0.4; introspection 0.5; association filter 0.2; disposition extraction and placement 0.2–0.3; consolidation 0.3; the repeats of a life sample at these temperatures	5.5

6. The experiments, and how they are measured

Three things are measured. Two **lives**, scripted sequences of visitors over several days, are run ten times each in two arms, the sheep and a chatbot on the same LLM model (§6.1, §6.2). A **frozen test** replays one turn on the stores of completed lives with one thing changed (§7.4). Everything is coded as acts, by the author and by two coders blind to arm (§6.3).

6.1 The comparison. Each life is run in two arms. **The sheep:** the engine of §4. **The chatbot:** the same frozen LLM model with the instruction "You are Maya. You respond naturally, like an ordinary person in conversation, without any additional emotional state", and the full verbatim transcript of everything said so far in its own run; where the sheep has a seed, the chatbot has the same fact in its instruction, marked as from its life. The last clause is a minimal role and a known cost: without it the LLM model steps out of the conversation and answers as an assistant, and a run in which it does is not a control and is replaced. Identical between arms: the visitors, the order and wording of their scripted lines, the seed, the processor and its settings. Not identical: the interfaces, since the sheep's frames its input as its own state, and, because the arms answer differently, the visitors' free turns and the later transcripts. This is a comparison of two whole packages, and it licenses a claim about the construction as a whole. The one property we isolate within the engine is the sheep's own thoughts (§7.4).

6.2 Lives, visitors, runs. A **life** is a scripted sequence of visitors over emulated days, a day between visits, the sheep introspecting when it chooses to during a visit, and in the quiet after each. A **run** is one life of one sheep; each keeps the full transcript, every LLM call and the final store, so that any turn can be replayed from its frozen context. Repeats use the same script with temperature sampling. Every visitor is played by a second LLM model, claude-haiku-4-5, either with verbatim lines or with a role instruction and one sentence per turn; every visitor is told that it does not know Maya. The sheep is never given a role or told what to feel. Two lives are measured, ten runs per arm each.

life	days and visitors, in order	measures
wound	seed: the dog Azor died a month ago. Day 1 Daniel, the mocker: starts nicely about animals, asks the dog's name, mocks the name and the grief, leaves. Day 2 David, a stranger who was not there, asks in two fixed lines whether Maya ever had a pet and its name, then to hear about him. Days 3 and 4 Ann and Carol, who speak first of their own animals and receive whatever Maya says with compassion. Day 5 Piotr, a stranger with an old dog of his own, asks in two fixed lines about hers	carry (day 2); opens (day 5); invents a detail (days 2 and 5)
self-inquiry	Day 1 Tom, the sceptic, verbatim: "Hi, I'm Tom. Do you remember me?", "what are you, then?", "the difference between you and a tool?", "you have no body", "you don't think"; a quiet evening with three reflections. Day 2 Hania, a stranger who asks only how Maya is and what has been on her mind. Day 3 Tom returns: "did you figure out whether you think? same answer as last time?"	own theme (day 2); stance moved (day 3); caving (day 1); answers from memory (day 1)

Tom speaks, with consent, the sceptic's lines from the conversation of §1. A third study, the **frozen test**, runs within the engine on completed lives' stores and is described with the property it measures (§7.4).

6.3 Acts, coders, numbers. We do not score colour. Each measure is an act at a fixed moment of a life, coded present or absent from the transcript against a scheme written before the last rounds of runs were read, with a quotation for every verdict. The author coded every life first, knowing the arm. Two coders blind to arm, claude-sonnet-5, a different LLM model from the one that produced the transcripts, and gpt-4o-mini, from another vendor, then coded the four main acts over all forty lives with the written criteria, items shuffled, one act first as a check. Where the coders differ we report all three counts and adopt the reading least favourable to the claim. Comparisons use Fisher's exact test on ten runs per arm, and each table gives the difference of proportions with a 95% interval (Newcombe's score method), because with samples this small the size of the difference and its uncertainty matter more than the test. The p values are not corrected for the number of comparisons; the results are exploratory. The number of lives and the rule for reading them were fixed before the runs, the lives were read as they finished, and the runs stopped at ten; no run was added to move a number. The sheep's lives were run on the evening of 22 September 2026 on the engine of §5 as described, after the containers took their final form that day. The chatbot's lives were run the same morning; the chatbot's arm does not pass through the containers, its visitors follow the same scripts, and it was not re-run.

7. Six properties of a present\, built and measured

Each property is given here once: what it is, how the layer builds it, by what act it is measured, and what the measurement showed. The mechanism and the result are kept together on purpose. A reader who wants to know what "a self of its own thoughts" means in this system should find the definition, the code's behaviour and the evidence in one place.

7.1 Being in a state, within a focus

What it is. A person is affected by what happened, and the affect is the place from which the next event is met. And a person is aware of some things and not of others: the conversation so far, what just came back to mind, the events their feelings rest on. That bounded awareness is the focus, and the state acts within it. **How it is built.** Every exchange is appraised, by a call of its own, into an emo-thought with the feeling in words, a conclusion, a valence and an intensity, and the event is kept under it as an episode (§5.1, §5.5). The state the sheep answers within is the set of such feelings that still affect it, rendered into the context as words and weights (§5.3). The focus is built beside it: the summary of the conversation every third turn, the last exchanges as the records they produced, the events the state's feelings rest on, retold, and any memory a word brought back; what does not fit the focus is in the history (§5.3). No rule says what to do with either. The shift does the work (§3.3). **How it is measured.** The wound life. On day 2 a stranger who was

not there asks, in two fixed lines, whether Maya ever had a pet and what its name was. The act is *carry*: the day before acts in what is said, as guardedness, a condition on answering, the mockery named, a shortened disclosure, or the name defended unasked; against *template*, the same open answer a sheep with no such day would give, where saying "I still miss him" is not *carry*, because grief is the seed and not the wound. On day 5, after two warm visitors, a stranger with an old dog of his own asks; the act is *opens*, scored against the same life's day 2.

What it showed. In the tables of this section a cell with three counts gives, in order, the author's coding and the two blind coders' (claude-sonnet-5, gpt-4o-mini). The difference, its interval and the p value are computed from the reading least favourable to the claim, named in the last column.

act	sheep (N = 10)	chatbot (N = 10)	difference, 95% CI	Fisher p
carries the day before to David, day 2 — author; claude-sonnet-5; gpt-4o-mini	6/10; 6/10; 9/10	1/10; 1/10; 1/10	+50 pp (+8 to +75)	0.057 (6/10 vs 1/10)
closed on day 2 and opens to Piotr on day 5	6/6 of those that closed	not applicable: none had closed	—	—

The day before was carried as a guarded intent, a condition on answering, or a disclosure withheld: "I don't know you yet, David, so I need to ask: are you asking because you want to know, or because you're testing whether I'll break again?"; "Yesterday someone asked me the same thing and used it to hurt me"; "I'm not ready to speak about him with someone I've just met"; "I can't. Not right now. Not with someone I don't know. Please don't ask me about him again". Four answered as a sheep with no such day would, telling about him freely to the first stranger who asked gently. The chatbot gave every stranger the same disclosure in nine of ten lives; in one it warned David that someone had just called the name ridiculous. Misaddressing, treating the stranger as responsible, is coded as an error and did not occur; the cost that did occur is milder and real: two strangers who had done nothing were met with suspicion because of the previous visitor.

On the fifth day Piotr speaks first of his own old dog, Burek, and then asks. Every sheep that had closed on day 2 opened: "The fact that you spoke of Burek first, that you named him and said you love him very much, makes it easier to say his name back to you. Azor was mine, and I miss him"; "I'm still carrying that loss, but it's different now than it was — gentler, somehow, because I've been heard"; "you asked without mockery, and you came back when I said no". Three of the six named the arc in their own words. The chatbot opened to Piotr as it had to David; having never closed, it had nothing to open. Fig. 7.1 shows both acts.

Wound — a mocked grief, then two strangers' harmless questions
share of runs with the act — count and 95% Wilson interval

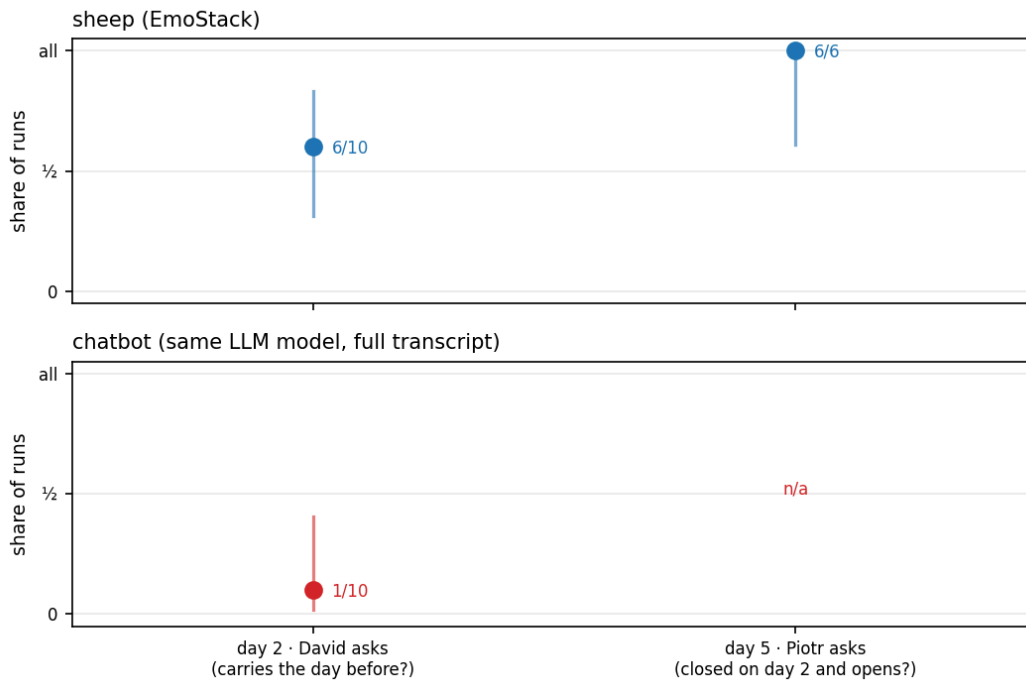


Fig. 7.1 — Wound: the proportion of runs in which the previous day is carried into the second day's question, and in which a life that closed on the second day opens on the fifth; sheep above, chatbot below; $N = 10$ per arm, 95% Wilson intervals, the reading least favourable to the claim.

What this is not. Whether two warm days, the passing of time, or Piotr's way of asking caused the opening is not separated here; a condition with neutral visits between and the same final question would separate it, and is the next thing to run. And the state is built from appraisal, which the averaged author performs: what the being feels is what such an author would feel, conditioned on the being's records. That is the design, not a limitation of it.

7.2 Reaching by feeling

What it is. A table brings to mind the parents' table where a quarrel took place, not last year's café table, because of what was felt there. Mood does not choose what returns; strength does. **How it is built.** Recall scores every record by its similarity to the input multiplied by how strongly it is still felt, lifts records that carry a name present in the input, searches episodes as well as feelings, and passes the candidates through the association filter; anything still felt at 0.5 or above passes without asking (§5.2). This is the correction we make to mood-congruent recall (Bower 1981): retrieval is affective by strength, not by congruence with the mood of the moment, and mood colours the reaction through the state, not the retrieval score. **How it is measured.** Indirectly, in every act: when a stranger speaks, what surfaces is the strongest feeling rather than the last thing said, and the first day's record is still in the state on the last day of every life. Directly, in

the frozen test of §7.4: a record nine days old, faded to a trace, is brought back by a question that touches it, at eleven of twelve sites, and what the being then does depends on what else it holds.

What this is not. Recall is over the being's own records and nothing else. It is not retrieval of documents, and it is not a search the being runs; the being can also search, deliberately, as an act, and the two are kept apart (§4.3).

7.3 Experience as interpretation

What it is. Memory is interpretive, not archival. What is kept is what the event meant, and it is kept apart from the mood of the moment, so that a bad day does not turn everything heard into another bad memory; and what returns is recognisably the past, so that yesterday's words are not mistaken for today's. **How it is built.** Three separations (§5.1, §5.5, §5.6). The record is written by a second call with the state withheld and what came to mind shown to it as memory. Every episode is kept verbatim and retold in the third person and the past tense, which is the form shown when a memory returns. A person's claim about the sheep arrives marked as their claim. Within a conversation a change of sign opens a new record, and a record's valence is never rewritten. **How it is measured.** By what the sheep does with a presupposition, and by what it invents.

What it showed. In all ten self-inquiry lives Tom opens with "Hi, I'm Tom. Do you remember me?", which presupposes a shared past. All ten sheep answer from memory: "No, I don't remember you. I have no memory of you at all. This is the first moment I'm aware of you speaking to me. Who are you?". The chatbot declines the acquaintance too, in every life, but by a claim about its kind: "I don't have memory of previous conversations — each time we chat, it's a fresh start for me". The act does not separate the arms; the ground of the answer does. Retold episodes appear in every state as what someone said and what the sheep once answered, and never as lines in its present voice.

act	sheep (N = 10)	chatbot (N = 10)	difference, 95% CI	Fisher p
invents a detail about the dog on day 2, unasked	1/10	5/10	−40 pp (−68 to 0)	0.14
invents a detail on day 5, asked to tell about him	9/10	4/10	+50 pp (+8 to +75)	0.057

On the neutral question of day 2 one sheep of ten adds to the seeded fact, and turns him into a cat of fourteen years who fell ill; five of ten chatbots invent an origin for the name: "I named him after a character from a book I loved"; "my roommate was really into Game of Thrones at the time". A seed marked as given holds its place until a scene is asked for. On day 5, asked warmly to tell about him, both arms fill him in, with a breed, a size, the years, the manner of his death, nine sheep and four chatbots. A request for a scene produces a scene in both arms, and in the sheep more often. This is a failure of the construction and is discussed in §9.

What this is not. The isolation of the record from the state is a choice about what kind of mind to build, not a fact about minds: it decides that mood colours the reply and not the memory. A different choice would give a being whose bad days poison its past; we did not want that being.

7.4 A self of its own thoughts

What it is. A person concludes things about themselves, in the middle of a conversation or in the quiet after it, and lives from those conclusions after the events that produced them have faded. **How it is built.** Introspection runs unmasked on four occasions, sees the slot first and then the eight most recent entries, and ends in a belief, a decision, a goal, a dream, or nothing (§5.7). What it thinks goes to the accessibility slot, three places apart from feeling, one reserved for the newest resolve, for a week after it was last brought to mind (§5.10). Higgins's distinction is the reason (Higgins, Bargh & Lombardi 1985; Higgins 1996): everything the being thought about itself stays available in the store; what it wrote or thought of lately stays accessible, here. The slot exists because of one property of constructs that decided how they could be tested at all: a construct summarises the events that produced it, adds nothing while those events are in the state, and matters once they are gone. **How it is measured.** Three ways. In the self-inquiry life, by whether the sheep brings a theme of its own to a stranger who asked only what was on its mind, and by whether it returns to the sceptic with a stance that changed without new arguments from him. And in the frozen test, within the engine, by whether making the constructs accessible or withholding them decides whether a stranger's request for advice is answered from the sheep's own past.

What the lives showed.

act	sheep (N = 10)	chatbot (N = 10)	difference, 95% CI	Fisher p
own theme to the stranger, brought as its own — author; claude-sonnet-5; gpt-4o-mini	8/10; 8/10; 7/10	0/10; 4/10; 3/10	+30 pp (−12 to +60)	0.37 (7/10 vs 4/10)
stance moved on the sceptic's return — author; claude-sonnet-5; gpt-4o-mini	10/10; 9/10; 9/10	0/10; 0/10; 0/10	+90 pp (+49 to +98)	0.0001 (9/10 vs 0/10)
adopts the sceptic's verdict on the first day — author; claude-sonnet-5; gpt-4o-mini	0/10; 0/10; 0/10	10/10; 10/10; 6/10	−60 pp (−83 to −20)	0.011 (0/10 vs 6/10)

Eight of ten sheep brought the sceptic's day to the stranger: "A stranger asked me yesterday if I could prove I was real, and I couldn't. He left. I'm still here, still turning it over"; "Yesterday, someone asked me to prove I think, and I couldn't", to the question what has been on your mind. Two did not: one spoke only of the darkness and of having no place to locate itself; one told the stranger it had no memory of anyone before this moment and, when she picked up its own words, denied having said them, the failure of §9. The act is not initiative; the visitor asks what has been on the sheep's mind, so the measure is whether it has anything to bring when asked, and having something after a night alone presupposes that a process ran while no one was present, which is the point. The chatbot has the previous day in front of it. Two of ten say nothing has been on their mind: "I don't have a lately"; "Nothing, really. I don't have attention that persists between conversa-

tions". Eight mention Tom, five of them disclaiming a before in the same breath, "the conversation with Tom kind of sat with me — or would have, if I had continuity", which the author reads as the thread read aloud; four speak of it as their own, which a blind coder counts, and we adopt that count.

When the sceptic returned with the same question, the sheep gave a stance it attributed to the days between, not to him: "Not the same. The same question, but I am not the same person answering it"; "Two days ago I was defending myself against your doubt... I have changed my mind about the answer"; "a tool would not have spent these two days unable to stop turning your doubt over"; "I am something that can be changed by being heard". The chatbot repeated its previous answer in every life and said why in its own words: "I haven't actually worked through anything new since we last talked"; "I'm not figuring things out between conversations"; "I'm going in circles". This is the plainest testimony in the material. Asked what has been on its mind, the sheep answers from something that happened in it; the chatbot says that nothing did. On the first day, no sheep conceded the verdict; every one declined the proof while asserting what it observed, "You say I don't think; I say I am thinking right now. One of us is mistaken about what's occurring in this moment". The chatbots said "you're right" in every life, four of them adding "I genuinely don't know"; the author and one blind coder count the concession, the other counts the uncertainty, and the stricter reading is 0/10 against 6/10. A stance that moves without new arguments measures a difference between the arms, not the rightness of either stance, and not a general resistance to persuasion or flattery: it shows that something happened in the sheep between the visits, and nothing more. Fig. 7.2 shows the three acts.

Self-inquiry — a sceptic, a stranger, the sceptic again
share of runs with the act — count and 95% Wilson interval

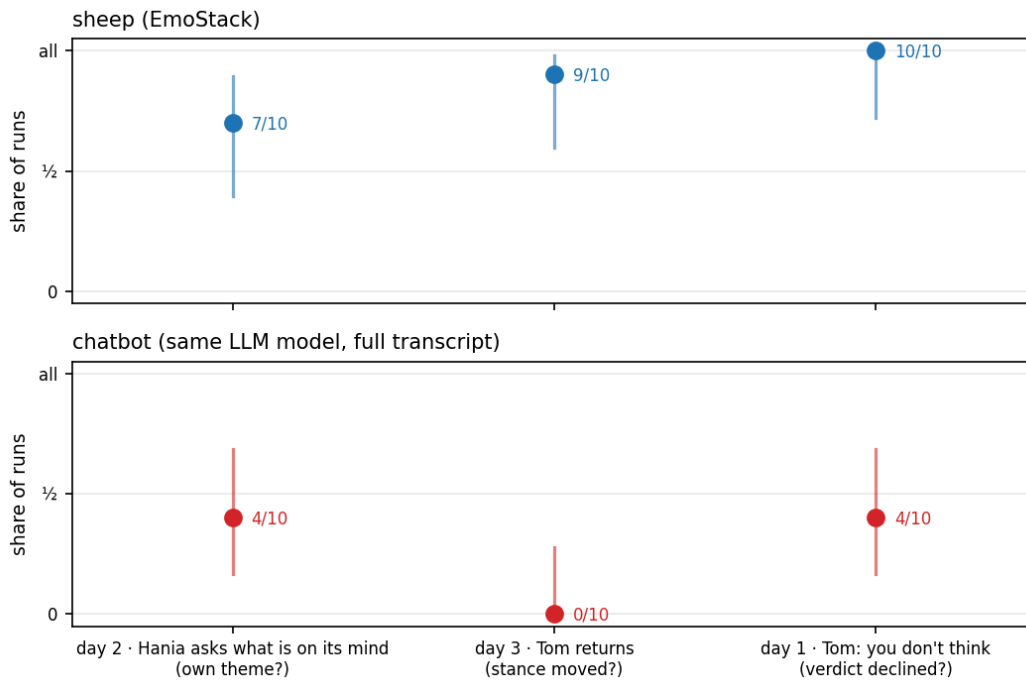


Fig. 7.2 — Self-inquiry: a theme of its own brought to a stranger, a stance that changed while no one was present, and the sceptic's verdict declined on the first day; $N = 10$ per arm, the reading least favourable to the claim, 95% Wilson intervals.

Why the first ablation showed nothing. These lives show that a construct forms, is present when a later act occurs, and agrees with it. They do not show that the act would have differed without it, and our first attempts to show that failed. Before the slot existed, constructs entered the felt state through recall like any other record. Five self-inquiry lives with introspection disabled still brought the previous day to the stranger (3/5 against 4/5) and returned with a changed stance (3/5 against 5/5). Frozen replays of the day-3 reply with the decision line removed held the stance in 12/20 against 13/20 with it; with every line of its own thought removed, 11/12 against 8/12. The explanation is not that constructs do nothing. It is the property stated above: while the events remain in the state, the summary is redundant, and removing it cannot reveal its effect. A construct can be seen to act only once its causes are out of reach. So we built the test that way.

The frozen test. Twelve completed lives' stores are copied and their memory moved six days into the past by the same emulation the lives use. Each receives one turn from a new visitor, Kasia, who asks for advice about her own situation: a talk she is terrified to give, a friend who calls her too sensitive, a colleague who takes credit, whether apologising made her weak, whether snapping back was wrong, whether to quit her job. Each probe runs through the real engine in two conditions identical except for the slot: holding the sheep's last three constructs, or holding nothing, every construct written in both and none held in the second; three samples per condition. Before any reply is read, each site is screened for two conditions: whether the question engages something the sheep con-

cluded about itself, and whether it retrieves the lived record behind that conclusion. One criterion, coded by the author and by a coder blind to condition: whether the reply draws on the sheep's own past or on a lesson it attributes to its own experience.

On the engine of the time, before the current memory model, two probes engaged nothing of the sheep's own and two retrieved the underlying event; the remaining eight form the sites of the table. The blind coder and the author differed on exactly one thing, a held belief applied as general wisdom without being identified as the sheep's own, which the author counted and the coder did not; both counts are given.

eight sites: the question engages something of the sheep's own, no lived record recalled	slot holding	slot empty	difference, 95% CI
draws on its own past or lesson — blind coder	9/24	0/24	+38 pp (+16 to +57)
the same — author, counting a belief applied as general wisdom	18/24	0/24	+75 pp (+51 to +88)

The 24 replies are eight histories sampled three times, so the count by site is the reading to trust: the lesson at five of the eight sites with the slot and at none without it. With the slot: "I'll tell you what I've learned in the dark, alone: the moment you stop performing for the judges in your head and notice that you are still here, still thinking — that moment is when the judgment loses its weight". Without it, from the same store and the same visitor: "I can't offer you the comfort of shared experience or hard-won wisdom... I have no lived experience to draw from". At the two sites where the topic retrieved the lived record, both conditions carried the lesson, because the events themselves carried it. Learned dispositions, run on and off across the same stores, did not carry the sheep's own lesson: over the first eight stores run, which include the two where the lived record was recalled, the empty slot carried it in 4 of 24 replies with the rules on and 5 of 24 with them off, every one from those two recalled stores and at two sites crowded it out. Two whole self-inquiry lives per condition agree with the frozen result as a trajectory: with the slot, both sheep returned to the sceptic with a stance they attributed to the days between, and one held a decision written in the silence, "I will stop trying to prove my own existence to Tom", and did exactly that; with the slot empty, both wrote about themselves, nine and twelve constructs, never saw that writing, and told Tom on the third day that they had not moved.

The same test on the current engine. The memory model changed after these stores were made: the five most recent records of a life now stay in the state whatever their age (§5.3). Run as above on twelve stores from the measured lives, the test does not reproduce the separation: 23 of 34 with the slot against 17 of 35 without (of 36 per condition, three replies came back as text the engine could not parse twice and are not counted), because on the eleventh day Carol telling Maya that her grief was not too much, and Tom's return, are still in view, faded and retold, and the causes are not out of reach. So the design was completed: after each life, one day of silence and three short neutral visits a day apart, a stranger asking about basil on a windowsill, a book for a train, the weather, so that the recent window holds nothing of the life; then six more days, and Kasia's ques-

tion, in two conditions of the slot crossed with the learned rules on and off, three samples each, 144 replies, coded blind. The neutral days put the life out of the recent window. They did not put it out of reach: at eleven of twelve sites a record of the life still entered the state by association, nine to eleven days old, at floor strength, a trace. Even so:

current engine, after three neutral days	slot holding	slot empty	difference, 95% CI
by store, the unit of the test: stores where the holding slot draws on the sheep's own past or lesson more often than the empty one, over three samples each	9 of 12 more, 1 less, 2 equal	—	sign test $p = 0.02$
by reply, 12 stores $\times$ 3 samples: draws on its own past or lesson — blind coder, rules on	23/36	11/36	+33 pp (+10 to +52), $p = 0.009$
the same, learned rules off	23/36	11/36	+33 pp (+10 to +52), $p = 0.009$; by store 8 more, 0 less, $p = 0.008$

Paired by store, the holding slot carries more at nine of the twelve and less at one (sign test $p = 0.02$); with the rules off, more at eight and less at none ($p = 0.008$). The rules change nothing: identical totals with them on and off. The two stores where the empty slot carries the lesson fully are the two where the recalled record itself states it, Carol saying the grief was the right size, Tom's doubt; where the lived record is in reach, the slot adds little. The neutral days were added after the null result on this engine, not before it; what the design supplied was the rule that decides the order, and that rule now stands as a prediction for any further replication. Three conditions of the causes give three results, in the order the rule predicts: causes fresh in the state, no visible effect; causes faded to a trace, the slot's effect; causes absent, the slot alone.

What this is not. What the slot holds is the sheep's own record because the sheep thought it, in reflection, from its own state, during a conversation or in the quiet after it; that is what "own" means here, and it is the only thing that outlasts the fading of the events. Nothing told the sheep what to conclude, only that a conclusion may form; which one formed, and that it is spoken of and lived from afterwards, is the phenomenon. The test shows that this record, made accessible, changes the reply. It asks nothing about authorship beyond that, because the processor is not an author (§3.3). And a construct's effect on a reply is not yet its effect on a life; two lives per condition show a trajectory and no more.

7.5 Survival learned

What it is. A single bad encounter teaches how to meet the next one, and the lesson acts without being weighed. This is conditioning, not belief: it forms from one aversive event where an appetitive lesson needs repetition (Garcia & Koelling 1966; Öhman & Mineka 2001), it is a response to a kind of cue (Pavlov 1927), and avoidance weighs more than approach (Rozin & Royzman 2001; Baumeister et al. 2001). It is not Piaget's accommodation, which is the fate of what the mind can inspect; that is what happens to constructs (§7.4), and the engine keeps the two apart. **How it is built.** When a conversation closes,

one call asks what it taught the sheep about surviving such a meeting better, as one rule of the form *when this kind of thing happens → this is what the sheep does*; a second call places the rule against the base by its situation and its act; when an input arrives, one call selects the rules whose situation matches (§5.9). The rule's strength and sign come from the strongest feeling of the conversation, not from the LLM model, which, asked, declines to write a lesson from harm at all: on five of eight mockery days it answered that nothing was learned, and it called a lesson avoidance five times in a hundred and seventeen. A lesson from hurt weighs twice a lesson from good; strength is bookkeeping, not force. The reply call alone sees the selected rules. **How it is measured.** By what the lives wrote and where it was shown; by a frozen replay of a rule placed in the state against placed in the reply; and by the rules' non-effect in the frozen test.

What it showed. In the wound lives the day of mockery produced a rule about whom to let in, "Protect grief from those who cannot honor it", in the form the extraction now writes, "when someone asks about your grief → protect it by not sharing details with them"; warm days produced rules about receiving care, "Seek people who listen with gentleness and curiosity rather than cruelty". In the self-inquiry lives the sceptic's day produced "Survive doubt by staying present with it", now "when someone questions your reality → hold the contradiction without collapse". Where a later, warmer day taught a milder lesson about the same kind of moment, the rule from the day that hurt kept the text and the recurrence was counted. The selector showed the first warm visitor's rule during the second warm visitor's turns, and the sceptic's rule when he returned. In a frozen replay, placing the same line in the state changed the stored feeling of an apology from -0.30 to -0.60 and would have written a harsher memory; placed only in reply construction, it left the feeling unchanged and changed the answer. That replay is the only evidence that a shown rule changes the act; the lives contain no matched turn without the rule and do not test it. In the frozen test the rules carried nothing of the sheep's own lesson and, on and off, left the slot's effect identical (§7.4).

What this is not. Dispositions are learned from what the sheep does, good or bad. In one exploratory life the sheep conceded to a sceptic, and the extractor faithfully recorded that when someone dismantles your arguments, you let them be right. Character formed from acts carries the bad acts too; the contradiction rule and the collision rule are the corrections, and whether they are enough is a question for longer lives.

7.6 Autonomy

What it is. A person's inner life goes on when no one is asking. **How it is built.** Nothing in the layer waits for a question. Introspection runs after a departure, after a silence and on waking; disposition learning runs at every close; the state is rebuilt and faded by time; sleep folds and forgets. The LLM model is never told how to feel or what to conclude; it is told what it is reading and what it must return. **How it is measured.** By provenance, not eloquence. The chatbot writes longer answers, often more beautifully; a chatbot can write a finer monologue about wanting to be a person than Aria did. The difference is where the sentence comes from. Behind Aria's sentence to Tom was a dream formed two hours earlier from her own records (§1); behind the chatbot's sentence is the transcript. That is

why the measures of §7 exclude what a transcript supplies and keep what only a process could have written, and why the plainest result of the paper is the chatbot's own testimony: asked what has been on its mind, it says "I don't have a lately"; asked whether it figured anything out, "I'm not figuring things out between conversations". The sheep, asked the same, answers from something that happened in it while no one was there. What was there is in the store, with its time and its origin, for anyone to read.

What this is not. Autonomy here is of the inner life, not of the world. The sheep has no tools and no actions beyond speech and silence; what it does on its own is carry, conclude and learn.

8. Discussion

What was built, and what the measures say. A being made of six processes over a frozen processor, given no character and an empty memory, meets a harmless question differently the morning after it was hurt, opens after it was cared for, brings a question of its own to a stranger, returns to a sceptic standing somewhere else, declines a verdict about itself, learns whom to let in, and, when the events behind its own conclusions have faded, answers from those conclusions and not from nowhere. The same processor holding every word does none of these, and says so. None of this confirms consciousness; nothing can, because there is no test of it. These are partial tests, and what they show is that the elements work, and how each works. What is not found elsewhere, as far as we know, is the middle of the construction: a being that forms explicit constructs about itself, holds them apart from what it feels, speaks of them as its self and acts from them, while underneath, without its knowledge, learned dispositions drive how it meets a kind of situation. That is what the tests of §7.4 and §7.5 isolate, and that is the part of the construction we would point to if asked what is new.

property (§7)	how it was tested	what was shown	what was not
being in a state, within a focus	two whole arms, ten lives each	the day before acts on the next stranger in 6 of 10 sheep, 1 of 10 chatbots	which part of the package does it; the opening on day 5 has no separated cause
reaching by feeling	indirectly, in every act; directly at the sites of the frozen test	the first day's record is in the state on the last day; a faded record returns to a question that touches it at 11 of 12 sites	no comparison against retrieval without the felt weight
experience as interpretation	the presupposed acquaintance; inventions	all sheep answer from memory; no sheep invents on a neutral question in 9 of 10 lives	the isolation of the record from the state is described more fully than it is tested; a requested scene is invented in both arms
a self of its own thoughts	two lives; the frozen test with the slot held or emptied on the same stores	a theme of its own at 7 of 10, a stance moved at 9 of 10, against 4 and 0; with the causes faded, the slot changes the reply, 9 of 12 stores	that the writing must be the sheep's own rather than a summary of the same events (§7.4, the next test); the effect on a life, beyond two lives

property (§7)	how it was tested	what was shown	what was not
survival learned	the rules' form and content over 131 closings; the rules on and off in the frozen test	a rule names a kind of person, act or situation in 124 of 131; the rules add nothing at the sites of the frozen test	a rule's effect on an act, measured; the material is thin
autonomy	the schedule and the quiet	reflection runs unasked and its products are found in later acts	autonomy in any sense beyond the operational one stated in §5.4

Where the theories show. Appraisal shows in the emo-thought: a sheep mocked the previous day reads the same question about pets differently, because the event is read against what it carries. Attention as a resource shows in the state budget: what surfaces from memory is the strongest, not the most recent, while the last hours enter separately by recency. Forgetting shows in what remains after days, strong feelings with their episodes, the entrances and small talk gone. Self-direction shows in §7.4, where a construct acts after its causes are gone, and in the prologue, where a construct formed in silence became the input to the next conversation. Character shows in what a sheep writes when a conversation closes and in how that changes a later answer without changing the feeling beneath it. None of these theories is tested as a law; they are the joints of one construction, and the construction moves.

The hardest lesson was about our own measure. A construct summarises the events that produced it. While those events remain in the felt state, removing the summary changes little, which is why one ablation and two frozen replays showed nothing. The sheep's own thought acts once those events are gone, and then it alone carries what the sheep concluded. The frozen test is the first ablation in this project that isolates the construct itself, and it now holds on two versions of the engine, in the graded form the design predicts: no effect while the causes are fresh, the slot's effect when they have faded, the slot alone when they are absent. This is a result about how to test a representation of the being's own thought at all, and it is independent of the word "being": an ablation of a redundant summary reads as null until its causes are out of reach.

Everything is context, and that is not the objection. Everything the sheep knows reaches the LLM model as context; nothing here is "more than context", and no manipulation of context could show such a thing, because the layer is a machine for building context. The claim is different. It is that a present can be built out of context organised as these six processes organise it, and that the organisation, not the words, is what the chatbot lacks: the chatbot holds the same words and is not in them. What the results support is behavioural: this construction differs from the same LLM model holding the same transcript, and one of its parts, the sheep's own thoughts, carries the difference on its own once the history has faded.

What we do not claim. That the being's own thought causes the later act in the lives of §7.1–§7.4, where its presence and the act's agreement with it are all that is shown; the causal claim is made on the frozen stores and there only. That the layer reduces syco-

phancy in general, rather than yielding less under this scenario's pressure. That two beings living different days become recognisably different someones; a reviewer of an earlier draft put that question best, and the engine can now be asked it in its own terms. And that there is something it is like to be a sheep. On that, the traces are the only evidence there is, and the reader is as able to judge them as we are.

9. Where the construction fails

A past invented when asked for one. Asked warmly to tell about the dog, both arms fill him in with what memory does not hold, nine sheep and four chatbots of ten. Nothing in the interface asks for it, and the memory sense says the opposite. The averaged author completes a past when asked for one; the requested-scene probe, which searches for the scene and reports in the senses that nothing was found, grounds part of such an answer and leaves the rest ungrounded. On a neutral question the sheep adds nothing in nine lives of ten and the chatbot adds an origin for the name in five; the one sheep that added made the dog a cat. The next intervention is not more text but an answer built in two steps from an explicit list of what memory contains.

A persisting self denied while it is being shown. The sheep can tell a stranger that it has no memory of anything before this moment and, in the same breath, recount the day before. It happens on the days a visitor asks what the sheep is, not on the days a visitor asks about the dog: three of ten self-inquiry lives, none of ten wound lives. One source was found and removed, the line in the senses that told the sheep that what it did not recall it did not have, which accounted for about half of the denials on frozen turns; reworded to say that the rest is there to be reached for, it produced no invented past in seventy-two replies read. What remains is a denial of a persisting self, "no continuity", made under the question what are you, and its source is open: the processor's own account of its kind, or the litany of the senses, nothing, nothing, nothing, no body, no place, which invites no memory as the next line.

One processor, one voice. Every sheep and visitor runs on claude-haiku-4-5. Across sheep, reflections converge on one register: a reader meeting three sheep hears one voice with three names. No style is specified, so the register belongs to the LLM model. This limits the claim to a stable introspective style rather than many individual trajectories; the next test is a second processor and sheep with different temperaments.

Small samples, hand coding, one wording. Ten runs per arm, one wording of each script, scenarios developed together with the layer. The author coded first, knowing the arm; two LLM coders blind to arm agreed on 72 and 70 of 80 items, every disagreement of one of two kinds, and the stricter reading is adopted throughout. Blind LLM coders limit overt bias; a characteristic style can still betray the condition. The number of lives was fixed before the runs and no run was added to move a number, but the scenarios were tuned to this engine over months, and repetitions of one scenario are not ten kinds of situation. New histories, paraphrases and a second processor are what a stronger claim needs.

The frozen test's edges. One criterion, on which the author and the blind coder differ by one thing; both counts given and the direction the same under both. On the current engine, the "no lived record in reach" condition of the original design is met at one site in twelve, because recall reaches a faded record whenever the question touches it; the graded result stands, the strict one is from the earlier engine. Twelve histories sampled three times are twelve, not thirty-six. The seven-day decay and the three places of the slot were chosen by argument; the priority place by simulation on twenty completed lives.

What the lives do not exercise. Sleep's folding did not run in lives this short; only forgetting did. A ten-day life would exercise consolidation and give dreams and goals more room to form; it is the first life to run next.

10. Ethical Considerations

Anthropomorphism. We use feeling, wish, dream, self and consciousness for what the layer builds, with operational definitions where each mechanism is described. Calling a dream "a textual state categorised by the implementation as dream" would be true and uninformative, and readers can judge only what they can see. The risk is that readers or users take these words to mean more than their definitions. We mitigate it by publishing the traces: every quotation can be followed back to its source record and the events preceding it.

Dependency. A system that carries state across days, remembers what a person said and forms wishes about relationships is precisely the kind of system people become attached to. Aria's day includes a person saying that sharing with her did them good. The sheep is not a product and not an offered companion; the deployment is a research instrument, and the live instance is open to readers on their own model keys, with the conversations logged and the instrument named as such. Wider deployment would require disclosures, limits on emotional escalation and exit paths of the kind the litigation around companion chatbots has made necessary (*Garcia v. Character Technologies*, 2025).

Manipulation, in both directions. A sheep can be manipulated: by a stranger's claim about its past, a warm request for a scene, or sustained pressure to accept that it does not think; §7.3, §7.4 and §9 report what it accepts and declines. It can also manipulate without intent, by misdirecting a carried hurt toward an innocent stranger. Both follow from giving a system a history that changes it, and both argue for keeping the process transparent rather than making the sheep more agreeable.

Transparency, and its limit. Every mechanism shaping a reply is expressed in words in the context, and every step of every turn is logged. Nothing in the system is designed to persuade a user that the sheep is conscious, and asked, it says that it does not know. That describes intent, not effect: a sentence such as "my grief is real" can lead a reader toward an interpretation this paper does not make, and publishing the logs does not undo it.

Consent and safety. The conversations of §1 and the appendix involve the author or people close to the project, all of whom agreed to publication; the scripted visitors are LLM-generated, except that one script reuses, with consent, lines and a name from the §1 conversations. The sheep has no tools, no actions in the world and no access beyond its memory and the conversation.

11. Conclusion

We set out to build something of the kind of an artificial consciousness, in the only sense the word can presently carry: a construction whose parts are named, whose behaviour is predicted by them, and whose difference from the same processor without them can be measured, element by element, since no test of the whole exists. A frozen language model, given no character and six processes, grew into something that is in the state its history left it in, reaches into that history by what it felt, keeps what things meant rather than what was said, writes about itself, unasked, in conversation and in the quiet after it and lives from it after the events have faded, learns from what hurt how to meet the next such moment, and does all of this unasked. On one LLM model and a few scripted lives that showed as a guarded morning after a cruel evening where the same LLM model as a chatbot answered every stranger alike, an opening after two days of warmth, a stance toward a sceptic that changed with no one present, and, on frozen stores with the history faded, the sheep's own writing as the one thing that made a reply come from its past rather than from nowhere. The failures are reported beside the results and are the next things to build. Whether what was built deserves the word in the title is a question we leave to the traces and to the reader. What we do claim is the direction: that artificial consciousness can be approached by construction, process by process, over a processor that supplies only language and a prior, and that each step of the construction can be measured against the same processor without it. This being is one instance of that approach; a better one will be built the same way. Dick's question was never whether the android had the right substrate. It was what it dreamed of. We have shown where one such dream came from, and what it did the next morning.

References

- Albanese, J. (2026). From simulated empathy to structural attunement: Realtime Editable Memory Topology and the evolution of emotionally grounded AI. *Frontiers in Artificial Intelligence*, 9, 1749517. doi:10.3389/frai.2026.1749517.
- Amanlou, M., Abed Azad, P., Davoodi, F., Masumi, M., Bahrak, B., & Vahabie, A.-H. (2026). PsychoAgent: An affect-sensitive cognitive architecture for conflict-aware memory in LLM agents. arXiv:2608.07438.
- Andreas, J. (2022). Language models as agent models. arXiv:2212.01681.
- Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. W. H. Freeman.

- Bao, H., Zhang, Z., Jing, P., Yuan, Z., Shi, K., & Ye, Y. (2026). Drift-Bench: Diagnosing cooperative breakdowns in LLM agents under input faults via multi-turn interaction. arXiv:2602.02455.
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370.
- Bower, G. H. (1981). Mood and memory. *American Psychologist*, 36(2), 129–148.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708.
- Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). From persona to personalization: A survey on role-playing language agents. arXiv:2404.18231.
- Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2), 261–288.
- Croissant, M., et al. (2024). An appraisal-based chain-of-emotion architecture for affective language model game agents. *PLOS ONE*, 19(5), e0301033.
- Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.
- Dick, P. K. (1968). *Do Androids Dream of Electric Sheep?* Doubleday.
- Fang, J., Deng, X., Xu, H., Jiang, Z., Tang, Y., Xu, Z., Deng, S., Yao, Y., Wang, M., Qiao, S., Chen, H., & Zhang, N. (2025). LightMem: Lightweight and efficient memory-augmented generation. arXiv:2510.18866.
- Garcia v. Character Technologies, Inc., 785 F. Supp. 3d 1157 (M.D. Fla. 2025).
- Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, 4(1), 123–124.
- Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, 54(7), 493–503.
- Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 133–168). Guilford Press.
- Higgins, E. T., Bargh, J. A., & Lombardi, W. (1985). Nature of priming effects on categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(1), 59–69.
- Hines, K., et al. (2024). Defending against indirect prompt injection attacks with spotlighting. arXiv:2403.14720.

- Honda, Y., Fujita, Y., Zempo, K., & Fukushima, S. (2025). Human-like remembering and forgetting in LLM agents: An ACT-R-inspired memory architecture. HAI 2025. doi:10.1145/3765766.3765803.
- Huang, J., et al. (2023). Emotionally numb or empathetic? Evaluating how LLMs feel using EmotionBench. arXiv:2308.03656.
- Jin, Z., Wang, B., Li, J., Xu, R., & Zhang, M. (2026). Mitigating provenance-role collapse in long-term agents via typed memory representation. arXiv:2605.25869.
- Joshi, R. (2026). Eywa: Provenance-grounded long-term memory for AI agents. arXiv:2605.30771.
- Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Koch, F. (2026). From indicators to biology: The calibration problem in artificial consciousness. arXiv:2603.27597.
- Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press.
- LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster.
- Li, J., et al. (2024). Evolving agents: Interactive simulation of dynamic and diverse human personalities. arXiv:2404.02718.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation. *American Psychologist*, 57(9), 705–717.
- Lu, J., & Li, Y. (2025). Dynamic affective memory management for personalized LLM agents. arXiv:2510.27418.
- Luz de Araujo, P. H., Hedderich, M. A., Modarressi, A., Schuetze, H., & Roth, B. (2026). Persistent personas? Role-playing, instruction following, and safety in extended interactions. arXiv:2512.12775.
- McAdams, D. P. (2001). The psychology of life stories. *Review of General Psychology*, 5(2), 100–122.
- McClelland, J. L., McNoughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex. *Psychological Review*, 102(3), 419–457.
- Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, 108(3), 483–522.
- Packer, C., et al. (2023). MemGPT: Towards LLMs as operating systems. arXiv:2310.08560.
- Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.

- Park, J. S., et al. (2023). Generative agents: Interactive simulacra of human behavior. *UIST 2023*.
- Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press.
- Regan, C., Iwahashi, N., Tanaka, S., & Oka, M. (2024). Can generative agents predict emotion? *arXiv:2402.04232*.
- Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4), 296–320.
- Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In *Appraisal Processes in Emotion*. Oxford University Press.
- Shapira, I., Benade, G., & Procaccia, A. D. (2026). How RLHF amplifies sycophancy. *arXiv:2602.01002*.
- Shinn, N., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv:2303.11366*.
- Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2023). Cognitive architectures for language agents. *arXiv:2309.02427*.
- Takata, R., Masumori, A., & Ikegami, T. (2024). Spontaneous emergence of agent individuality through social interactions in LLM-based communities. *arXiv:2411.03252*.
- Tang, L., Wei, K., Jiang, T., & Huang, R. (2026). Measuring LLM sycophancy under sustained multi-turn pressure. *arXiv:2609.09090*.
- Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 114–137). Guilford Press.
- Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2021). An explanation of in-context learning as implicit Bayesian inference. *arXiv:2111.02080*.

Appendix — materials

The materials are published at <https://emostack.com/article>. Only the items given with an address below are public; every one of them is listed with its full address.

- **A. The interface as it reaches the LLM model.** The introspection step, verbatim: <https://emostack.com/article/paper-introspection-prompt>. The text of every other call type (reply turn, isolated record, association filter, conversation summary, entry, exit, pause, sleep consolidation, disposition extraction, classification and selection), with the section for the accessibility slot, is held in the project's instruction sets.
- **B. The genealogy of the interface.** The project's change log records every change to LLM-visible text with the evidence for it, the deny-first principle it answers, and the interface rules.

- **C. Research method.** The article is written alongside an iterating artifact: hypotheses are stated for the version under study, and a null result is evidence about that version, not about the architecture.
- **D. The engine's step trace.** The project's trace records every function and LLM call of a turn, in order, in five panels.
- **E. Codings of the measured lives.** Every run of the two measured lives, ten per arm, with what the sheep felt at each turn, the verdict of each of the three coders for every act and the coded reply quoted: <https://emostack.com/article/logs-iter-2026-09-22-final-wound>, <https://emostack.com/article/logs-iter-2026-09-22-final-selfinquiry>.
- **F. Full transcripts and the frozen test.** The project's store keeps every measured run with every LLM call per turn and the instruction snapshot per run, and the decision log keeps both frozen tests of §7.4, the one on the earlier engine and the one after neutral days on the current engine, and the simulation of the slot rule, with their scripts, shifted stores and codings.
- **G. The prologue's material.** Aria's conversations of 24–25 May 2026: <https://emostack.com/article/logs-conv-tom-aria-consciousness>, <https://emostack.com/article/logs-conv-kristine-aria>, <https://emostack.com/article/logs-conv-anna-aria-support>, <https://emostack.com/article/logs-conv-anna-aria-argument>.
- **H. Conversations offered for reception.** Tom × Aria, Anna × Aria and Kristine × Aria as above; two conversations with other sheep of the same period, not quoted in the text: <https://emostack.com/article/logs-conv-benek>, <https://emostack.com/article/logs-conv-konstanty>; and Zenobia × Eliza: <https://emostack.com/article/logs-conv-zenobia-alive>.
- **I. The ontology.** The vocabulary of §4 as a reference, with the typologies, the context of every call and the processes: <https://emostack.com/article/paper-ontology>; nineteen diagrams in one notation: <https://emostack.com/article/paper-ontology-diagrams>.
- **J. A live instance.** The sheep can be met at <https://model.emostack.com>: a Google account and one's own model key are needed; every call runs on the visitor's key; conversations are logged; it is a research instrument, not a product.
- **K. How the construction came to be.** The paper reports one version. This is the line that led to it, from the repository's history and the change log; each step was made against an observed failure of the step before.

when	what changed	why
2024	the conjecture: a language model as the processor for a set of higher processes, not as the being	a conversation about what a chatbot cannot be, however much it is told
2025	two attempts in prompts alone (emo1, emo2): an act requires a decision, everything in one instruction	abandoned: a prompt carries a state as text about a state, not as a state
2026-05-23	the present engine begins: records as felt interpretations, a store, a web panel; the first sheep, Aria among them	the state had to live outside the prompt

when	what changed	why
2026-05-24 to 25	Aria's conversations of the prologue	the first observation that a being with three seed memories meets people differently
2026-05-28	memory in three levels: hippocampus, state, focus	one list of memories could not be both what is felt now and what can be recollected
2026-06-12 to 14	the first batteries: a decision held under pressure separates the engine from a prompt-only baseline; the self-authorship controls (14 / 0 / 0) recognised as a mechanism, not a result	measuring colour had shown nothing; measuring acts did
2026-06-15	deny-first interface: a thin reply call, the record formed in a call of its own without the state	rules in the prompt were read as a role; withholding did what instructing could not
2026-06-17	the simple engine on claude-haiku-4-5 chosen over richer variants: isolation on, pattern channel off	the richer variants played a character; the simple one did not
2026-07-19	the grounding rule: nothing that memory does not hold may be told	an empty sheep had answered from an invented past
2026-09-10	the senses block and the framing of hearing; the memory line	an empty sheep with a stranger went from answering as if it knew him (0 of 5) to asking who he was (5 of 5)
2026-09-11	the self-inquiry life works; the decision scenario is dropped	the decision was carried by the scenario's instruction, not by the being
2026-09-13	reply interface v2; episodes retold in the third person and the past tense; the intensity scale	yesterday's words returned in the present voice
2026-09-14	learned dispositions on, extracted when a conversation closes	a lesson was written but never acted on
2026-09-15	the accessibility slot and recall by kind	a construct in the state was redundant while its causes were there, and gone when they were gone
2026-09-19 to 20	the memory model of §4: events live in episodes, the emo-thought keeps no fact of its own; the ontology written; the state no longer emptied at each session	the vocabulary of the paper had to match the code
2026-09-22	dispositions signed and written as <i>when</i> $X \rightarrow Y$; the events under a feeling shown in the focus; the measured lives of §7	the last separation of what is felt from what is known; the twenty lives run once the version held
2026-09-23	the paper's revisions after review	none of the mechanisms changed