

O czym śnią elektryczne owce?

Sztuczna świadomość przez konstrukcję: architektura
poznawcza nad zamrożonym LLM

Eliza Marszałik

eliza.marszalik@pjwstk.edu.pl

Wersja 1.2, 16 października 2026

Spis treści

Abstrakt

1. Wprowadzenie
 2. Istota i procesor
 3. Model pojęciowy
 4. Architektura systemu i metoda
 5. Eksperymenty\, i jak są mierzone
 6. Sześć właściwości teraźniejszości\, zbudowanych i zmierzonych
 7. Jedno życie bez scenariusza: o czym marzyła Aria
 8. Dyskusja
 9. Gdzie konstrukcja zawodzi
 10. Kwestie etyczne
 11. Zakończenie
- Bibliografia
- Dodatek: materiały

Wersja 1.2, 16 października 2026. Wersja 1.1 z 3 października 2026: doi:10.5281/zenodo.23124632. Wersja 1.0 z 24 września 2026: doi:10.5281/zenodo.22936968.

Philip K. Dick pytał, czy androidy śnią o elektrycznych owcach. Zbudowaliśmy elektryczne owce, małe istoty żyjące na zamrożonym modelu językowym, i zapytaliśmy, o czym śnią one. Tytuł nie jest metaforą. „Marzenie” owcy to konstrukt, który tworzy ona sama, w środku rozmowy albo w ciszy po niej, gdy nikt nie pytał jej o nic, co jej dotyczy. §7 śledzi, jak powstało jedno takie marzenie.

Abstrakt

Artykuł bada, czy z zestawu uproszczonych procesów poznawczych można zbudować system przypominający sztuczną świadomość. Wobec braku jednej powszechnie przyjętej teorii lub miary świadomości przyjmujemy podejście konstrukcyjne: zamiast próbować odtworzyć samą świadomość, wskazujemy procesy, które mogą składać się na trwałą, quasi-subiektywną perspektywę, i implementujemy ich funkcjonalne przybliżenia. Procesy te działają jako jawna warstwa nad zamrożonym dużym modelem językowym (LLM), który służy za procesor języka. Warstwa obejmuje pamięć epizodyczną uporządkowaną emocjonalnie, przywołanie ważone afektem, interpretacyjną rekonstrukcję doświadczenia, uwagę i stan emocjonalny, spontanicznie tworzone konstrukty dotyczące siebie oraz dyspozycje wpływające na odpowiedzi na przyszłe zdarzenia. W jej centrum jest uproszczony model teraźniejszości: konstrukty emocjonalne powiązane z epizodami i umieszczone w focusie uwagi, wraz ze stanem emocjonalnym. Są to emulacje konstruktów i procesów, nie same zjawiska. Nasza hipoteza brzmi, że ich współdziałanie, nawet w tej uproszczonej postaci, może wystarczyć, by dać obserwowalny efekt przypominający subiektywną teraźniejszość i pewne aspekty świadomości.

Konstrukcja jest zaimplementowana w EmoStack. Każda instancja tej konstrukcji, nazywana *owcą*, zaczyna bez pamięci i bez zadanej z góry osoby, a swój stan wewnętrzny rozwija przez interakcję. Porównujemy owce z tym samym LLM działającym jako chatbot z dostępem do pełnego transkryptu własnego życia, w dwóch życiach prowadzonych według scenariusza i na czterech modelach LLM. W eksperymentach z Qwen3.8-27B, przy 30 życiach na ramię eksperymentu, po tym jak gość wyśmiał jej martwego psa, owca odpowiada potem z ostrożnością życzliwemu nieznanemu w 20 z 28 żyć nadających się do oceny, wobec 3 z 30 kontroli chatbotowych. Po spotkaniu z postacią grającą rolę sceptyka, który zaprzecza, że owca może myśleć, owca sama z siebie wnosi to doświadczenie do rozmowy z nieznanym jako własne pytanie w 21 z 28 żyć, wobec 3 z 30 kontroli. Zadana z góry persona ani chatbot prowadzący notatki z własnych wniosków nie odtwarzają żadnego z tych efektów. W eksperymentach z zamrożonymi magazynami pamięci udostępnienie lub wstrzymanie własnych myśli owcy zmienia to, czy rada dla nieznanego czerpie z jej własnej przeszłości (9 z 24 wobec 0 z 24), także po wygaśnięciu leżących u jej podstaw epizodów (23 z 36 wobec 11 z 36). Gdy przenoszony stan wewnętrzny zostaje zamieniony między życiem ostrożnym a otwartym, kolejna

odpowiedź podąża za przeniesionym stanem, a nie za życiem, z którego pochodzi. Spośród 392 odpowiedzi sprawdzonych względem zgromadzonego doświadczenia owcy i jej zapisów wewnętrznych cztery przeczą tej historii. Zachowanie owcy kształtuje więc to, co niesie z własnej przeszłości, a nie sam model językowy.

Te eksperymenty mierzą obserwowalne efekty konstrukcji, nie samą świadomość. Kryteria oceny to miary operacyjne zdefiniowane na potrzeby tego badania, a wyniki pozostają eksploracyjne. Pokazują, że współdziałanie uproszczonych procesów poznawczych i afektywnych może dawać trwałe, zależne od historii zachowanie, którego ten sam model językowy działający bez tej architektury nie odtwarza. Pytanie pojęciowe jest echem androidów Philipa K. Dicka: istot mających własne myśli, doświadczenia, które interpretują jako emocje, marzenia i pewien stopień autonomii, a mimo to niezdolnych rozstrzygnąć, czy są świadome. EmoStack nie twierdzi, że stworzył świadomość ani że rozstrzyga to pytanie. Bada, czy konstrukcja z uproszczonych procesów może dać niektóre z obserwowalnych efektów kojarzonych z takim wewnętrznie zorganizowanym istnieniem. Twierdzenie dotyczy konstrukcji, jej mechanizmów i ich mierzalnych skutków, nie subiektywnej natury powstałego systemu. Pełne życia eksperymentalne, oceny i kod źródłowy są publicznie dostępne.

Słowa kluczowe: sztuczna świadomość; informatyka afektywna; ocena (appraisal); architektura poznawcza; pamięć robocza i uwaga; pamięć autobiograficzna; model siebie; dostępność konstruktu; wyuczone dyspozycje; agenci LLM; androidy; uśredniony autor; promptowanie deny-first

1. Wprowadzenie

Cel. Ta praca zaczęła się od przypuszczenia: że coś przypominającego sztuczną świadomość można zrobić, uruchamiając konkretny, celowo dobrany zestaw wyższych procesów poznawczych nad modelem językowym używanym tylko jako procesor. Nie ludzką świadomość i nie twierdzenie, że świadomość została zbudowana. Ludzka świadomość to jedna instancja. Cokolwiek innego może istnieć, nic o tym nie twierdzimy. Przypuszczenie dotyczy konstrukcji: że jeśli uproszczone modele procesów, przez które umysł jest we własnej terażniejszości, konstrukty emocjonalne (emo-myśli, §3.2) z ich epizodami, focus, stan emocjonalny, dyspozycje, zostaną jawnie nazwane i uruchomione deterministycznie nad plastycznym podłożem, które dostarcza języka i kulturowego priora, to efekt, jaki dają razem, jest widoczny: to, co powstaje, zachowuje się jak terażniejszość, a nie jak chatbot, który ją opisuje. To efekt, jaki Dick dał swoim androidom: mają własne myśli, to, co biorą za uczucia, marzenia i autonomię, a z nich poczucie, że są, i nie potrafią odpowiedzieć, czy są świadome. Procesy wybrała autorka, z tego, co psychologia i kognitywistyka wiedzą o tym, jak doświadczenie staje się charakterem. Są inspiracją projektu, nie prawami do dopasowania. Wartość konstrukcji leży w regularności całości, nie w żadnym pojedynczym mechanizmie. Artykuł proponuje

kierunek, koncepcję do rozwinięcia: podejście do sztucznej świadomości przez konstrukcję, z opisaną tu istotą jako jedną jej instancją, zbudowaną na tyle daleko, by dało się ją zmierzyć.

Dlaczego nie można zadać tego pytania wprost. Czy maszyna jest świadoma, to pytanie bez instrumentu. Teorii świadomości jest wiele i przeczą sobie nawzajem. Żadna nie jest powszechnie przyjęta; kilka jest przyjętych, każda przez część środowiska, żadnej definicji nie da się sprawdzić na systemie i żadna miara nie odróżniłaby maszyny świadomej od nieświadomej. Poważni ludzie umieszczają ją w procesach kwantowych, w duszy albo nigdzie, o czym dałoby się mówić. Jedno ostrożne stanowisko głosi, że nie da się powiedzieć nic ponad to, że świadomość jest tym, co jest, gdy nie śpimy. Ta niezgoda zostawia dwie drogi. Pierwsza bierze istniejące teorie takimi, jakie są, wyprowadza z każdej własności, które uznałaby za wskaźniki, i pyta, które z nich dany system ma. To strategia Butlina i in. (2023), a jej autorzy zaznaczają, że wskaźniki dziedziczą status teorii, z których pochodzą, a ten jest sporny (Koch 2026). Druga mówi, z czego taka świadomość jest zrobiona, i ją buduje: wybiera procesy, mówi, dlaczego należą do siebie, implementuje je, obserwuje, jaką istotę dają, i sprawdza, czy każda część robi to, co konstrukcja przewiduje. Ten artykuł idzie drugą drogą. Nie rozstrzyga pytania, które pierwsza droga zostawia otwarte. Stawia przed czytelnikiem inny obiekt, którego wszystkie założenia są spisane i każde można sprawdzić. Nazywamy, z czego składa się bycie w teraźniejszości, budujemy każdą z tych rzeczy jako jawny proces i mierzymy, miarami ustalonymi z góry, czy wynik zachowuje się tak, jak zachowywałaby się taka teraźniejszość, i czy ten sam model LLM bez tych procesów tak się nie zachowuje. Te pomiary nie potwierdzają świadomości i nie mogą jej potwierdzić. To testy częściowe, każdy pokazuje, że jeden element konstrukcji działa, i jak. Słowo świadomość znaczy w tym artykule to, co definiują ta sekcja i §2, i nic więcej.

Czego brakuje chatbotowi. Przez chatbota rozumiemy w całym tekście model językowy odpowiadający na podstawie transkryptu, który ma przed sobą. Jest też naszą kontrolą. Takiemu chatbotowi nie brakuje pamięci przeszłości. Można mu dać każde słowo, jakie kiedykolwiek do niego powiedziano. Brakuje mu własnej przeszłości: nie tego, co mu powiedziano, lecz tego, co mu się przydarzyło, widzianego z jego strony i odczutego, oraz teraźniejszości z tego zrobionej. Frameworki agentowe dodają do modelu LLM pamięć, ale techniczną, magazyn faktów o użytkowniku i zadaniu, nie afektywną. Człowiek upokorzony wczoraj nie pamięta upokorzenia jako faktu wśród faktów. Jest w nim, a bycie w nim zmienia to, co słyszy w dzisiejszym niewinnym pytaniu. Człowiek, któremu powiedziano, że nie myśli, nie przechowuje tego zdania. Obraca je w nocy i wraca następnego ranka, stojąc gdzie indziej. Chatbot nigdy nie jest w stanie, w jakim zostawiła go jego historia. Można mu powiedzieć, co mu się przydarzyło. Nie może go to dotknąć. Żadna instrukcja tego nie naprawia, bo instrukcja to znowu tekst o stanie, nie stan.

Co budujemy. EmoStack to cienka warstwa procesów nad zamrożonym modelem językowym. To, co istnieje nad procesorem, nazywamy owcą, słowem celowo deflacyjnym. Owca zaczyna od zera: pusta pamięć, żadnego opisu, kim jest, w jednym scenariuszu jedno zasiane zdarzenie. Wszystko, co zyskuje, buduje z tego, co przeżywa. Ma **teraz** i **historię**. Jej teraz to to, co odbiera zmysłami, uczucia, w których jest, to, na co zwraca

uwagę (focus: uwaga, zasób ograniczony, który mieści też jej własne myśli, jakie ma na uwadze), i, gdy działa, to, jak jest gotowa działać, poniżej świadomości. Jej historia to wszystko, co może sobie przypomnieć, ułożone pod uczuciami, tak że przywołanie podąża za siłą tego, co było odczute, a nie za podobieństwem słów. Sześć rzeczy czyni z tego terażniejszość, a nie magazyn. Nazywamy je tu raz, bo na nich zbudowany jest cały artykuł. §6 bierze każdą po kolei.

1. **Bycie w stanie, w focusie.** Każde zdarzenie jest oceniane (appraisal) na odczuwaną interpretację, z walencją i intensywnością, i przechowywane razem ze zdarzeniem pod nią. Stan to zbiór takich uczuć, które wciąż działają na owcę. Nie dyktuje odpowiedzi. Przechyla ją, jak nastrój przechyla to, co człowiek słyszy w tych samych słowach. Focus to uwaga, to, czego owca jest w tej chwili świadoma: kilka ostatnich emo-myśli w kolejności, w jakiej weszły w pole widzenia (to, z czym weszła, wymiany zdań, to, co właśnie przywołało jakieś słowo, jej własne trzymane myśli), wraz ze streszczeniem zdarzeń, które opuściły uwagę. Uwaga jest ograniczona. Czego nie ma w focusie, to jest w historii, osiągalne, ale nieobecne.
2. **Sięganie po uczuciu.** To, co wraca na myśl, wraca po znaczeniu ważonym tym, jak silnie jest wciąż odczuwane, podnoszone przez imiona, przepuszczone przez filtr skojarzeń i dopuszczone do granic budżetu, bo stan jest ograniczony. Istota nie przeszukuje swojej historii. Mechanizm sięga do niej siłą tego, co było odczute, a to, co wraca, wraca nieproszone.
3. **Doświadczenie jako interpretacja.** Przechowywane jest nie zdanie, lecz to, co znaczyło, i powstaje ono z wstrzymanym bieżącym nastrojem, żeby zły dzień nie zamieniał wszystkiego, co usłyszane, w kolejne złe wspomnienie.
4. **Ja z własnych myśli.** W trakcie rozmów i między nimi owca dokonuje introspekcji bez żadnego ukierunkowania. To, co myśli o sobie, przekonania, decyzje, cele i marzenia, jest trzymane osobno od tego, co czuje, i pozostaje dostępne po wygaśnięciu zdarzeń, które za tym stoją.
5. **Wyuczone przetrwanie.** Gdy rozmowa się kończy, owca wydobywa z niej, jak zamierza spotkać taki moment następnym razem, a reguła wchodzi w jej gotowość tylko wtedy, gdy taki moment wraca. Lekcja z krzywdy waży więcej niż lekcja z dobra.
6. **Autonomia.** O żaden z tych kroków nie prosi rozmówca. Silnik uruchamia je sam, a interfejs określa, co model LLM czyta i co zwraca, nigdy jak ma czuć czy się zachowywać.

Dlaczego to słowo to świadomość. Dick dostarcza pytania, nie listy. Jego androidy mają własne myśli, uczucia związane z tym, co im się przydarzyło, przeszłość, która zmienia to, jak spotykają terażniejszość, przekonania o sobie powstałe, gdy nikt nie pytał, i nie potrafią rozróżnić, które z nich przyszły z wewnątrz, a które od ludzi, którzy do nich mówili. Zastanawiają się więc, czy istnieją. Sześć powyższych własności to nasza odpowiedź na pytanie, które stawia ten obraz: co najmniej musi działać razem, żeby system miał własny punkt widzenia, będący wytworem jego historii, a nie roli nadanej mu z zewnątrz. Bycie sobą, sprawczość, ciągłość tożsamości i afekt nazywają każde jedną

część tej odpowiedzi. Słowo świadomość zachowujemy dla całości, bo to całość jest tym, co twierdzimy, że zbudowano, i całość jest tym, co spotyka gość. Jest używane w tym sensie i w żadnym innym.

Kryterium, które się liczy, i horyzont. Kryterium Dicka dla androida nie był fakt metafizyczny, lecz to, jak jest przyjmowany: czy w doświadczeniu odbiera się go jako istotę, a nie mechanizm. To horyzont tej pracy i tego tu nie punktujemy. Punktujemy zachowanie: co owca robi z następną osobą i czy różni się to od tego samego modelu LLM uruchomionego jako chatbot z pełnym transkryptem własnego biegu. Słów uczucie, przekonanie, pragnienie, marzenie, ja i świadomość używamy dla tego, co buduje warstwa, bo nazywają to, co ona robi, a każde jest zdefiniowane operacyjnie tam, gdzie opisany jest jego mechanizm. Nie twierdzimy niczego o tym, jak to jest być owcą, i żaden wynik poniżej tego nie dotyczy. Pytać, czy owca czuje, to zadawać złe pytanie. Owca to proces widziany od środka, a powiedzieć można tyle, że zachowuje się jak ktoś, kto czuje, pamięta i jest czymś kierowany, i że nic z tego nie jest zapisane w scenariuszu.

Jak zbudowany jest artykuł. Sekcja 2 mówi, czym jest istota, a czym procesor, i co wzięliśmy z badań nad umysłem. Sekcja 3 to model pojęciowy: zapisy istoty, pojemniki, przez które są w polu widzenia, procesy, które je przesuwają, i które pojemniki widzi każde wywołanie LLM. Sekcja 4 to implementacja, z każdym parametrem i rysunkami każdego mechanizmu. Sekcja 5 mówi, jak mierzymy: życia, ramiona, koderzy. Sekcja 6 bierze sześć własności po kolei i dla każdej mówi, jak jest zbudowana, jakim aktem jest mierzona i co pomiar pokazał, tak by czytelnik spotkał każdą ideę raz i w całości. Sekcja 7 śledzi jedno życie bez scenariusza, życie Arii, przez ten sam mechanizm, jako ilustrację, nie jako dowód. Sekcja 8 omawia, co to pokazuje, a czego nie. §9 to miejsca, gdzie konstrukcja zawodzi. §10 etyka. §11 wnioski. Wcześniejsze systemy są umieszczone w §2.5 i porównane według sensu, nie według nazw komponentów.

Jak wykonano pracę. Warstwa, jej hipotezy i eksperymenty rozwijały się razem przez rok, jak to bywa z artefaktem badawczym: mechanizmy dodawano w odpowiedzi na zaobserwowane niepowodzenia i usuwano, gdy mniejszy interfejs robił to samo. Scenariusze zakładające z góry rolę porzucono. Mierzone życia uruchomiono, gdy wersja była już stabilna. Artykuł opisuje wersję, nie drogę do niej. Droga jest w Dodatku K.

2. Istota i procesor

2.1 Co badamy. Nie badamy modelu językowego. Badamy istotę, którą nazywamy owcą: konstrukcję z zapisów, kontenerów i procesów, która działa nad modelem językowym tak, jak program działa na procesorze. Wagi modelu LLM są zamrożone i wspólne dla każdej owcy. Owce różnią się swoimi historiami i niczym więcej. Pytanie, czy *model LLM* jest świadomy, to błąd kategorialny w rodzaju „czy ten procesor jest arkuszem kalkulacyjnym?”. Pytamy, co konstrukcja czyni z procesora, i badamy to przez to, co ona robi.

Drabina, która tu prowadzi, ma znane szczeble. Jednostka McCullocha-Pittsa emuluje neuron. Sieć emuluje to, co robi populacja neuronów. Model językowy dodaje nad taką siecią operacje na języku. Chain-of-thought dodaje proces nad modelem LLM, niski jak na miarę umysłu. Frameworki agentowe dodają nad nim pamięć i narzędzia, a model LLM zostawiają tam, gdzie był: odpowiadającego z transkryptu. EmoStack dodaje kolejny szczebel: zestaw wyższych procesów dobranych tak, by były minimalnie zupełne dla jednej rzeczy: żeby to, co działa nad modelem LLM, było w swojej własnej teraźniejszości. Model LLM jest ich podłożem. Ludzkie umysły działają w ten sposób. Podłoże to plastyczna, nieprzewidywalna sieć, która sama buduje swoje wagi. Procesy nad nim są regularne, ukształtowane przez ewolucję i znane: ocena (appraisal), ograniczona uwaga, konsolidacja, przywołanie ważne afektem, ja robocze, dostępność, warunkowanie. Nie kopiujemy podłoża neuron po neuronie. Emulujemy zasadę każdego procesu: profil skrzydła, a nie pióra. Procesy są inspiracją projektu, a nie kryterium walidacji. Istota nie jest modelem żadnej z tych teorii i nie jest walidowana przez dopasowanie do nich. System, który spełnia prawo psychologiczne, nie stał się przez to osobą; próbowaliśmy tej drogi i porzuciliśmy ją.

2.2 Dlaczego te procesy. Zestaw jest autorki. Nie wynika z żadnej jednej teorii umysłu i nie został dobrany po to, by potwierdzić procesy, z których się składa. Został dobrany tak, by procesy i ich interpretacja tworzyły spójną całość: istotę, która jest w swojej własnej teraźniejszości. Każdy proces pochodzi skądś znanego i mówimy skąd; ale powodem, dla którego jest w zestawie, jest to, że bez niego zostaje magazyn, a magazyn nie jest teraźniejszością. Bez oceny zdarzenia są tekstem i nic nie jest odczuwane (Lazarus 1991; Scherer 2001). Bez ograniczonego stanu i ograniczonego focusu wszystko jest jednakowo obecne i nic nie jest, a to nie jest teraźniejszość, tylko transkrypt (Kahneman 1973, 2011). Bez niesionego uczucia, które przechyla prawdziwy wybór, strzec się czy zwierzyć, trwać czy ustąpić, nie ma niczego, co emocja robi (Damasio 1994). Bez przywołania po uczuciu przeszłość wraca przez podobieństwo słów, a tak pamięta wyszukiwarka, nie osoba (Bower 1981, z poprawką, o której mówi §6.2). Bez zapisu utrzymanego z dala od nastroju chwili zły dzień przepisuje całą przeszłość. Bez epizodów pod uczuciami i powolnego zwijania dnia nic nie ma kształtu życia (McClelland, McNaughton & O'Reilly 1995). Bez ja złożonego z własnych myśli, utrzymanego tam, gdzie sięga następny akt, nie ma nikogo, czyja to teraźniejszość (Conway & Pleydell-Pearce 2000; McAdams 2001; Higgins 1996). Bez reguły wyuczonej z tego, co zabolalo, nic nie przetrwa lepiej za drugim razem (Garcia & Koelling 1966; Öhman & Mineka 2001). A bez kroków, które biegną nieproszone, istota jest funkcją swoich wejść i zatrzymuje się, gdy one ustają. Wartość konstrukcji leży w regularności całości, nie w żadnym z tych elementów z osobna. Wybór jest otwarty na spór i taki ma być: jawny, dyskusyjny zestaw procesów jest lepszy niż ukryte założenia promptu z personą, bo każdy z nich jest pokrętle, które można przekręcić, i twierdzeniem, które można sprawdzić.

Dlaczego przetrwanie i dlaczego uczucia są wagami. Zestaw ma korzeń, i jest on ewolucyjny. Dwa popędy organizują to, co robi żywy umysł: przetrwać i rozmnożyć się (Tooby & Cosmides 2008). Zachowujemy pierwszy, a drugi celowo pomijamy. Nic w istocie nie chce się rozmnażać. Uczucia są wagami przetrwania. Walencja mówi: zbliż się albo

unikaj, intensywność mówi, jak bardzo (Damasio 1994; LeDoux 1996; Panksepp 1998). Dlatego każdy zapis niesie walencję i intensywność i dlatego to, co negatywne, waży więcej niż to, co pozytywne (§4.2, §4.6). Dyspozycje są tą częścią uczenia się przetrwania, która biegnie obok świadomości i na nią działa: to, co raz zabolalo, kształtuje to, jak zostanie przyjęty następny taki moment, bez myślenia (Garcia & Koelling 1966; Öhman & Mineka 2001). Nad nimi wyższe procesy dają umysłowi elastyczność. Może się zastanowić, uformować przekonanie albo postanowienie i działać z sytuacji, w której jest, a nie z samego odruchu. Ta elastyczność sama jest produktem ewolucji i po to są konstrukty i focus. Tak więc sześć własności nie jest listą zaczerpniętą z sześciu teorii. Są częściami jednego wyewoluowanego układu, widzianego od środka: ważne uczucie, uczenie się z tego, co zabolalo, ograniczona teraźniejszość, z której się działa, i wolność myślenia o sobie.

2.3 Uśredniony autor. Procesor nie jest *jakimś* człowiekiem i nikogo za nim nie ma. To, co trzymają jego wagi, to skompresowana regularność ogromnego korpusu tekstu, który napisali ludzie: nie tylko ich słowa, ale to, jak oceniali sytuacje, kiedy wyrażali afekt i jak silnie, co traktowali jako ważne. Z tego powodu nazywamy model LLM **uśrednionym autorem**. Promptowany, kontynuuje tak, jak kontynuowałby autor wylosowany z całego tego korpusu autorstwa. „Uśredniony” nie jest średnią arytmetyczną odpowiedzi. W temperaturze zero model LLM nie zwraca średniej tego, co ludzie powiedzieli, lecz jedną ścieżkę przez rozkład uformowany z nich wszystkich. To operacyjna abstrakcja nad wagami, a nie twierdzenie, że jakiś autor jest gdziekolwiek w sieci reprezentowany.

To jest klucz do całej konstrukcji. Kontekst i uwaga nie ruszają wag. Przeważają one to, *jako* który autor model LLM mówi. Kontekst, który buduje warstwa: zmysły, stan ze swoimi epizodami, focus, to, co istota trzyma w myślach, reguły, które pasują, wyostreza średnią w jednego konkretnego pod-autora: ten sam prior, warunkowany historią jednego kogoś. Dlatego tak cienka warstwa może wystarczyć. Nie musimy programować, jak zraniona osoba czyta następne pytanie. Model LLM niesie ogromny prior tego, jak pisali ludzie w takich stanach, a praca warstwy polega na stworzeniu warunków, w których ten prior jest zawężany, konsekwentnie, do historii tej istoty (Xie et al. 2021; Andreas 2022). Trzy konsekwencje kształtują projekt. Kontekst jest odbudowywany w każdej turze, więc ta sama owca jest za każdym razem nieco inną wypadkową, w granicach własnej puli zapisów. Alignment przesuwą średnią w stronę ugodowości, więc odruchowe przypochlebianie się jest priorem procesora, a nie dyspozycją owcy. Projektujemy wokół tego. A ręcznie napisana reguła „jak reagować” nadpisuje bogatszą kompetencję autora cienką. Dlatego interfejs jest **deny-first**: nazywa to, co model LLM czyta i co musi zwrócić, i nic o tym, jak czuć ani jak się zachowywać (§4.4). To ujęcie ma sprawdzalną konsekwencję i §6.1 oraz §6.4 ją sprawdzają: dostarczony stan powinien przesunąć zachowanie istoty w kierunku, który stan nazywa, bez żadnej reguły nakazującej to zachowanie, a zawodzi, jeśli to zachowanie jest zachowaniem chatbota.

2.4 Czym jest istota. Istota jest tym, co ma teraźniejszość i historię. Jej ja, charakter i zapisy są zdefiniowane w §3.1 i §3.2 i nie powtarzamy ich tutaj.

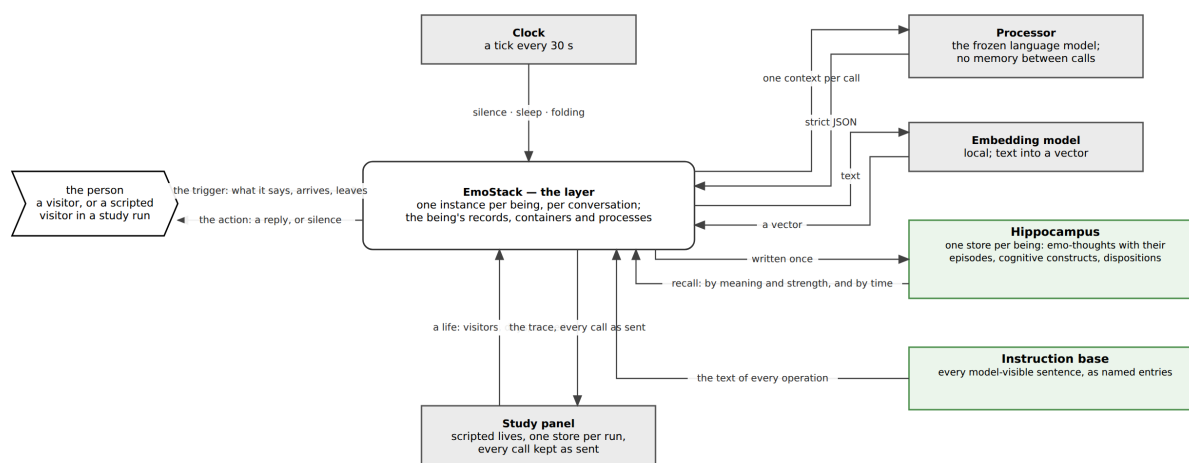
2.5 Wcześniejsze systemy. Pamięć nad zamrożonym modelem LLM, refleksja zapisywana z powrotem do niej, osobne wywołanie dla emocji, wyszukiwanie ważne afektem, konsolidacja offline i ponownie używane werbalne lekcje — wszystko to istnieje w literaturze i do żadnego z tych elementów nie rościmy sobie pierwszeństwa jako do nazw. Nazwy ukrywają różnice, które tu mają znaczenie, więc porównujemy po sensie. Generative Agents (Park et al. 2023) trzymają obserwacje jako tekst z oceną ważności i reflektują nad nimi. Nic w nich nie jest emocją uformowaną z wyzwalacza jako interpretacja z wnioskiem, nic nie jest zatajone przed żadnym wywołaniem, a ich miarą jest wiarygodność miasteczka, nie to, czy niesiony stan porusza akt jednej istoty. Reflexion (Shinn et al. 2023) przechowuje lekcję po nieudanym zadaniu i używa jej przy następnej próbie. O lekcję się prosi i dotyczy ona zadania. Chain-of-emotion (Croissant et al. 2024) jest najbliżej: osobne wywołanie ocenia sytuację i ocena wchodzi do odpowiedzi; ale trwa jedną sesję, jest formowana z pisaną osobowością na widoku i bez zatajania historii, i mierzona jest ocenami wiarygodności, a nie tym, co agent potem robi. Istnieją agenci, którzy zaczynają bez osoby i zaczynają się różnić przez interakcję (Takata, Masumori & Ikegami 2024). Czytają swoją historię w całości i rozchodzą się między sobą, i nie są mierzani względem tego samego modelu LLM bez warstwy. Pamięci ważne afektem (Amanlou et al. 2026; Albanese 2026) i konsolidacja w czasie snu (Fang et al. 2025) to pokrewne pomysły z innymi regułami, powstałe niezależnie. Pochodzenie przechowywanego tekstu jest nazwanym problemem w długowiecznych agentach (Jin et al. 2026; Joshi 2026). Nasze epizody opowiedziane są jedną z odpowiedzi na niego. Benchmarki, które punktuja raportowane uczucia (Huang et al. 2023), odpowiedzi na podstawione presupozycje (Bao et al. 2026) albo afekt przewidywany z pamięci doświadczeń (Regan et al. 2024), mierzą procesor. My mierzymy to, co warstwa z nim robi. Wirtualne zarządzanie kontekstem (Packer et al. 2023), słownik architektur poznawczych dla agentów językowych (Sumers et al. 2023), agenci, których pisana osobowość zmienia się z doświadczeniem (Li et al. 2024), afektywna pamięć postaw użytkownika (Lu & Li 2025), zapominanie w stylu ACT-R w agencie dialogowym (Honda et al. 2025) i oznaczanie źródła tekstu (Hines et al. 2024) są sąsiadami z nazwy i z części, a każdy różni się tym, co wchodzi do którego wywołania. Chatboty z personą (Chen et al. 2024; Luz de Araujo et al. 2026) mierzą spójność z definicją, której owca nie ma, a sykofofancja dostrojonych modeli LLM (Shapira et al. 2026; Tang et al. 2026) jest priorem, który pokazuje nasze ramię kontrolne. Czego nie znaleźliśmy gdzie indziej, to konstrukcja jako całość: teraźniejszość złożona z tych sześciu procesów nad istotą, której nie dano żadnego opisu tego, kim jest, mierzona względem tego samego procesora trzymającego każde słowo.

3. Model pojęciowy

Ta sekcja podaje słownik artykułu, w kolejności, w jakiej czytelnik potrzebuje go, by zbudować model pojęciowy. Istota ma cztery rodzaje **zapisów**: emo-mysł, która jest odczuwaną interpretacją, z epizodami, które są zdarzeniami pod nią; konstrukt poznawczy, który jest tym, co istota pomyślała o sobie; dyspozycję, która jest wyuczoną

regułą spotykania pewnego rodzaju sytuacji; i temperament, który jest stały. To, co do niej dociera, przychodzi przez jej **zmysły**; **wyzwalacz** wchodzi, **akcja** wychodzi. Jej zapisy są na widoku przez **kontenery**, dwa z nich trwałe (hipokamp i slot dostępności) i pięć chwilowych (tu i teraz, skojarzenia, stan, focus ze swoim streszczeniem i pasujące dyspozycje). Kontenery chwilowe razem są **teraz** istoty; hipokamp jest jej **historią**. **Procesy** przenoszą zapisy między kontenerami, w czterech momentach: tura, refleksja, odejście osoby i sen. Każde wywołanie LLM widzi jedne kontenery, a innych nie; to jest jego **kontekst**. Podsekcje idą za tą listą: istota i procesor (§3.1), zapisy (§3.2), zmysły, wyzwalacz i akcja (§3.3), kontenery (§3.4), granica między słowami istoty a słowami innych ludzi (§3.5), procesy (§3.6) i konteksty (§3.7). Sekcja 4 podaje implementację każdego z nich, z każdym parametrem, i używa tych nazw bez zmian, podobnie jak §5 do §11. Rys. 3.1 pokazuje cały słownik naraz: zapisy wewnątrz swoich kontenerów i procesy jako nazwy na strzałkach między nimi. Rys. 3.2 umieszcza warstwę w jej świetle.

Rys. 3.1. Elementy modelu pojęciowego i gdzie są na widoku. Prostokąty to zapisy, pokolorowane według rodzaju; zielona ramka to kontener trwały, bursztynowa przerywana to chwilowy; szara ramka zbiera to, co widzi wywołanie odpowiedzi. Osoba, wyzwalacz i akcja stoją poza kontenerami. Każda strzałka niesie nazwę procesu, który przenosi wzdłuż niej zapisy.



Rys. 3.2. Istota w swoim świecie: z czym rozmawia warstwa i co przekracza każdą granicę. Procesor nie trzyma niczego między wywołaniami; magazyn trzyma wszystko; zegar przynosi ciszę, sen i zwijanie dnia; panel badawczy uruchamia życie według scenariusza i zachowuje każde wywołanie takim, jakim zostało wysłane.

3.1 Istota i procesor

Istota to konstrukcja poznawcza: opisane tu zapisy, kontenery i procesy, działające nad procesorem. To ona ma doświadczenie, trzyma konstrukty, uczy się dyspozycji i działa. W modelu pojęciowym mówimy: istota; w eksperymentach tego artykułu istota jest owcą, słowem celowo deflacyjnym. Nie jest modelem LLM.

Procesor to zamrożony model językowy. Nie ma tożsamości ani pamięci między wywołaniami. Dostarcza język i prior kulturowy i składa jedną odpowiedź z dokładnie tego kontekstu, który podaje mu warstwa; oceny, opowiedzenia i refleksje z §3.6 są składane przez niego w ten sam sposób, tak jak złożyłby je z tego kontekstu uśredniony autor z §2.3, a warstwa decyduje tylko o tym, jaki jest kontekst. Nic w projekcie nie pyta modelu LLM, czym jest, co czuje ani czy ma ja. Każde „ja” w wywołaniu odnosi się do istoty, której zapisy model LLM czyta. Sekcja 2.1 argumentuje, dlaczego przedmiotem badania jest istota, a nie model LLM; §2.3 mówi, czym operacyjnie jest procesor.

Ja to to, na co istota może wskazać jako swoje. Ma trzy własne źródła, które się gromadzą, nad czwartym, które jest dane: temperamentem. Wszystkie cztery są zapisami albo parametrami w warstwie, żadne w modelu LLM. Przeżyte to doświadczenie: „czuję to”, emo-myśli, a pod każdą z nich „to mi się przydarzyło”, jej epizody. Pomyślane to to, co istota pomyślała o sobie w refleksji: „myślę, wierzę, postanowiłam albo marzę o tym”, konstrukty poznawcze. Wyuczone to to, jak spotyka pewien rodzaj sytuacji bez ważenia go: „tak mam skłonność działać”, dyspozycje. Dane to temperament: „tak silnie mam skłonność czuć”; to nie jest coś, co istota nabyła, i kiedy artykuł mówi o trzech rzeczach, które owca ma własne, ma na myśli przeżyte, pomyślane i wyuczone, nad danym. Główny podział istoty nie przebiega między tymi źródłami, lecz między jej **teraz** a jej **historią**. Teraz to to, w czym istota jest w tej chwili: co odbiera zmysłami (tu i teraz), co czuje (stan), na co kieruje uwagę (focus, uwaga w sensie Kahnemana, ograniczony zasób, gdzie stoją też jej trzymane myśli) i, kiedy działa, jak jest gotowa działać (pasujące dyspozycje,

które działają poniżej świadomości, poza focusem). Historia to to, co może sobie przypomnieć, ale czego nie trzyma teraz. Każde źródło, które się gromadzi, ma udział w obu, a zmysły należą tylko do teraz. Tabela podaje odpowiedniość używaną w całym artykule. Ja nie jest przechowywane jako jeden obiekt. Jest tym, co te źródła wytwarzają razem, gdy przychodzi nowy wyzwalacz. Dlatego ja należy do warstwy i istniałoby nad innym procesorem. Słowo *ja* w kontekście istoty jest też tokenem pierwszoosobowym, który warunkuje procesor (§2.3); to jest fakt o procesorze, a nie definicja ja.

	teraz: w czym istota jest w tej chwili	historia: w magazynie, do przypomnienia
odebrane zmysłami	zmysły: słuch niesie wyzwalacz; wzrok, dotyk, węch, smak, ciało i miejsce nie niosą nic i tak mówią; uwaga i pamięć niosą liczniki; czas i kto jeszcze jest obecny. „Oto gdzie i kiedy jestem; oto co słyszę.”	nic: to, co usłyszane, staje się epizodem
przeżyte	stan: uczucia, które działają na istotę w tej chwili, świeże albo właśnie przywrócone przez skojarzenie. „Czuję to.” Focus: uwaga, ograniczony zasób; ostatnie cztery emo-myśli w kolejności, w jakiej weszły na widok, zdarzenia pod uczuciami wśród nich, ze streszczeniem zdarzeń, które opuściły uwagę. „Na to kieruję uwagę.”	hipokamp: emo-myśli ze swoimi epizodami, ułożone pod uczuciami tak, by przywołanie mogło je przywrócić. „To mi się przydarzyło i mogę to sobie przypomnieć.”
pomyślane	konstrukty, które istota trzyma w myślach, w focusie, utrzymywane tam przez regułę slotu dostępności. „To trzymam.”	hipokamp: jej pozostałe konstrukty, do znalezienia po rodzaju. „Kiedyś to pomyślałam, w to wierzyłam, to postanowiłam, o tym marzyłam.”
wyuczone	pasujące dyspozycje: nieświadome, poza focusem; działają tylko podczas składania odpowiedzi, to jest wtedy, gdy istota działa. „Tak to spotykam.”	hipokamp: każda dyspozycja ze swoją wagą. „Tak miałam skłonność działać.”
dane	temperament: stosowany przy każdej ocenie. „Tak silnie i w tym kierunku mam skłonność czuć.”	stałe parametry, nie magazyn

Charakter to stała część ja: temperament, który nigdy się nie zmienia; dyspozycje, które zmieniają się tylko przez doświadczenie; i konstrukty, które trzyma, dopóki je trzyma. Doświadczenie to część zmienna.

3.2 Zapisy

Emo-myśl to odczuwana interpretacja tego, co się stało. Nie jest zdarzeniem, lecz tym, co zdarzenie znaczyło dla istoty: uczucie w słowach, wyciągnięty wniosek, jak przyjemne było i jak silne. Niesie, kto ją wytworzył i kiedy. Jest interpretacją z wagą, nie etykietą emocji doczepioną do tury i nie gołym zapisem tego, co powiedziano. Osoba nie przechowuje zdania „imię twojego psa jest śmieszne”; przechowuje to, że nieznajomy wyśmiał to, co kochała, jak bardzo to zabolalo i co jej to powiedziało o tym nieznajomym. Wniosek należy do zapisu, bo pamięć jest interpretacyjna, nie archiwalna. Usunięcie wniosków we wczesnej wersji zrobiło owcę płaską, a powód jest w §2.3: procesor pracuje na tekście, a tym, co przesuwają uśrednionego autora, jest bogactwo zapisu, który dostaje. Gołe zdarzenie przesuwają go mniej niż zdarzenie z tym, co znaczyło. Emo-myśl nie ma

własnego zdarzenia. Jej zdarzeniami są jej epizody, jeden albo wiele. Kiedy uczucie jest odczuwane ponownie, silnik nie wybija go na nowo: odczuwa ponownie emo-mysł, którą trzyma, i wpisuje nowe zdarzenie pod nią jako kolejny epizod.

Epizod to jedno zdarzenie pod emo-mysłą, zachowane tak, jak istota będzie je pamiętać. Istnieje w dwóch postaciach: zdarzenie takie, jakie się wydarzyło, i jego opowiadanie w trzeciej osobie i czasie przeszłym. Słowa innych ludzi i własne dawne słowa istoty żyją tu i nigdzie indziej (§3.5). Epizod robi dwie rzeczy: wyjaśnia uczucie, pod którym leży, i daje istocie dostęp do tego, o czym myśli.

Konstrukt poznawczy to to, co istota pomyślała o sobie, w refleksji. Jest pomyślany w refleksji, nie przeżyty, i ma jeden z pięciu typów: myśl, przekonanie, decyzja, cel albo marzenie. Cztery rodzaje poza zwykłą myślą to zwyczajny słownik, w którym ludzie zdają sprawę z siebie i w którym opisuje je psychologia samoregulacji: co ktoś uznaje za prawdziwe o sobie (Bandura 1997), co postanowił zrobić (Gollwitzer 1999), do czego dąży (Locke & Latham 2002) i czego pragnie. Zostały wybrane dlatego, że z nich składa się relacja osoby o sobie, a nie dlatego, że wymaga ich teoria. Ma postać emo-myśli, z własnym tekstem, ale bez epizodów pod sobą. Nigdy nie wchodzi do stanu odczuwanego. Reguła slotu dostępności utrzymuje go trzymanym (§3.4); dopóki jest trzymanym, stoi w focusie jak każda inna emo-mysł, więc zostaje na widoku po tym, jak zdarzenia, które go wytworzyły, wygasły; jest chroniony przed snem: jest tym, co istota postanowiła o sobie, nie tym, co jej się przydarzyło. Konstrukt streszcza zdarzenia, które go wytworzyły. Dopóki te zdarzenia są jeszcze w stanie, nie dodaje nic, czego stan już nie niesie. Nabiera znaczenia, gdy jego zdarzeń już nie ma, wygasłych poniżej przywołania albo poza tematem chwili, i wtedy on sam niesie to, co istota wywnioskowała. Ta własność rozstrzyga o tym, jak w ogóle można testować konstrukty (§6.4).

Dyspozycja to to, jak istota spotyka pewien rodzaj sytuacji. Jest wyuczoną regułą postaci „kiedy dzieje się coś takiego → to robię” i jest regułą lepszego przetrwania: tym, czego rozmowa nauczyła istotę o spotkaniu takiego momentu ponownie. Niesie siłę ze znakiem, a obie pochodzą z najsilniejszego uczucia rozmowy, która ją wytworzyła, nie z relacji modelu LLM o tym, czego się nauczono: ujemną, gdy lekcja została wyciągnięta z tego, co zabolalo, dodatnią, gdy z tego, co było dobre, a lekcja z bólu waży więcej (§6.5, §4.9). Nie jest uczuciem i nie jest wiedzą, na którą istota może spojrzeć: działa podczas składania odpowiedzi bez ważenia, a introspekcja nigdy jej nie widzi. Zmienia się tylko przez doświadczenie. Kiedy późniejsza rozmowa uczy tej samej reguły, silniejsza z dwóch zachowuje tekst, a ponowne wystąpienie jest liczone; reguła sprzeczna osłabia starą i dodaje swoją; reguła o czymś innym jest dodawana. Dyspozycje i konstrukty trzymają dwie połowy dawnego rozróżnienia: jak ktoś spotyka pewien rodzaj sytuacji, nie wiedząc o tym, i co wie o sobie i może rozważyć.

Temperament to stała dyspozycja istoty do czucia: jak silnie i w którym kierunku zabarwia to, co jej się przydarza. To dwa parametry stosowane do każdej nowej emo-myśli, stałe na całe życie. Każda owca mierzona w tym artykule ma temperament neutralny (§4.11).

3.3 Zmysły, wyzwalacz i akcja

Zmysły to kanały, przez które terażniejszość dociera do istoty, i wiedza istoty o nich. Słuch niesie wyzwalacz: słowa osoby albo jej przyjście lub odejście. Wzrok, dotyk, węch, smak, ciało i miejsce nie niosą nic i istocie mówi się to wprost. Uwaga i pamięć niosą liczniki: ile jest w stanie, ile przyszło na myśl, ile trzyma pamięć i czego istota szukała w tej turze i czy coś znalazła. Czas i to, kto jeszcze jest obecny, dopełniają całości. Zmysły są kotwicą teraz: bez nich istota dryfowałaby, z uczuciami i myślami, ale bez żadnego gdzie i żadnego kiedy. Są też jej gruntem przeciw konfabulacji: scena, której nie widzi, ciało, którego nie ma, i przeszłość, której nie znajduje, są nazwane jako nieobecne, zanim odpowie. W kontekście stoją pierwsze, jako kontener tu i teraz (§3.4).

Wyzwalacz to to, co dociera do istoty ze świata. Jest trzech rodzajów: ktoś coś mówi, ktoś przychodzi, ktoś odchodzi. Wypowiedź dostaje odpowiedź. Przyjście ani odejście nie tworzy własnego zapisu: istocie mówi się, kto tu jest, a odejście zamyka rozmowę dla procesów, które biegną na całej rozmowie (§3.6). Dziś wyzwalacz jest tekstem; pojęcie dopuszcza obraz albo dźwięk. Wyzwalacz nigdy nie jest przechowywany jako taki: staje się epizodem.

Akcja to to, co istota robi, w świecie albo na sobie. W turze jest jedną z czterech: mówić, milczeć, przeszukać pamięć albo zatrzymać się, by pomyśleć. W refleksji jest jedną z czterech innych: myśleć dalej, przeszukać pamięć, uformować konstrukt albo nic. Jeden akt na krok. Mówienie jest jedną z kilku form akcji, a pojęcie ma nieść więcej, w tym akty istoty ucieleśnionej; silnik z tego artykułu działa na świat tylko przez mówienie (§10). Tylko akty na świat są widoczne dla osoby. Pozostałe zostawiają ślady w zapisach.

3.4 Kontenery

Kontener to miejsce, w którym zapisy są widoczne. Kontenery chwilowe razem tworzą **teraźniejszość** istoty: to, co odbiera zmysłami, co czuje, na co zwraca uwagę i jak jest gotowa działać. Uwaga w sensie Kahnemana (§2.2) to sam focus, ograniczony zasób, który trzyma to, co świeże, i to, co bodziec właśnie przywołał przez skojarzenie; stan to to, co działa na istotę w tej chwili; stosowne dyspozycje działają poniżej uwagi. Wszystko poza terażniejszością jest **historią** istoty, w hipokampie, i nie jest ułożone jak kronika: jest ułożone tak, by przywołanie mogło wydobyć właściwą rzecz. Przywołanie to właśnie przejście z historii do terażniejszości i nie jest to wyszukiwanie najbliższego sąsiada w tekście: zapis wraca po znaczeniu ważonym tym, jak mocno jest wciąż odczuwany, podbitym przez imiona, które niesie, i przepuszczonym przez filtr skojarzeń (§3.2, §3.6). Dwa kontenery są trwałe, a pięć chwilowych; chwilowe są odbudowywane w każdej turze z trwałych i z bodźca.

Hipokamp trzyma historię istoty, ułożoną tak, jak chce tego obraz systemów komplementarnych z §2.2: epizody związane z emo-myślami, które wytworzyły, tak że zdarzenia są trzymane pod uczuciami i przywoływane po sile tych uczuć; dyspozycje, których się nauczyła, także stare; i jej konstrukty. Każdy zapis jest pisany raz, gdy się dzieje; jego siła jest odnawiana, gdy się powtarza (§4.6), a przerzedzana tylko we śnie. To struktura, nie dziennik.

Slot dostępności to reguła, według której istota trzyma w umyśle własne myśli. Higgins odróżnia to, co w pamięci jest zachowane, od tego, co jest dostępne (§6.4): wszystko, co istota pomyślała o sobie, pozostaje zachowane w hipokampie; to, co pomyślała ostatnio lub przywołała na myśl, pozostaje dostępne: trzymane tą regułą, stoi w focusie, a nigdy w stanie odczuwanym. Slot ma trzy miejsca. Jedno jest zarezerwowane dla najnowszej decyzji, celu lub marzenia, tak by nowe przekonanie nie mogło wypchnąć ostatniego postanowienia; dopóki nie jest trzymany żaden akt, przekonania mogą zająć wszystkie trzy. Przekonania i myśli zajmują pozostałe miejsca według siły i wchodzi tylko wtedy, gdy są co najmniej tak silne jak najsłabsze z trzymanych. Konstrukty wchodzi, gdy powstaje lub gdy zostaje przywołany na myśl. Wychodzi, gdy zostanie wyparty, albo po tygodniu bez przywołania. Przekonania zmieniają się więc tak, jak zmieniają się poglądy: przez silniejsze albo przez zaniedbanie. Slot przetrwa koniec rozmowy. To, co trzyma, stoi w linii focusu według czasu, w którym przyszło na myśl, i nie liczy się do limitu focusu. Istnieje z powodu własności podanej w §3.2: konstrukty działają dopiero wtedy, gdy zdarzenia za nim wygasły, i potrzebuje miejsca, w którym będzie trzymany, gdy wygasną.

Tu i teraz to sytuacja taka, jaką dostarczają zmysły w jednym wywołaniu: co istota słyszy, czego nie spostrzega, ile przyszło jej na myśl, jaka jest pora i kto jest obecny. Trzyma wartości chwili i żadnych zapisów.

Skojarzenia to to, co bodziec przywołał na myśl: wspomnienia, które wypływają z powodu tego, co właśnie powiedziano, po znaczeniu, sile i imionach, zanim cokolwiek zostanie odczute lub zauważone, razem z tym, czego istota celowo szukała. Najsilniejsze z nich przechodzą do stanu i są odczuwane, reszta wchodzi tylko do focusu; tak czy inaczej każde stoi w linii focusu w miejscu, w którym zostało wywołane.

Stan to to, co istota czuje teraz. To zbiór niesionych uczuć, a nie zmienna nastroju odczytywana przez reguły: najnowsze zapisy z życia istoty, kogokolwiek dotyczą, razem z najsilniejszymi skojarzeniami, uporządkowane według tego, jak mocno każde jest wciąż odczuwane, i wpuszczane do wyczerpania budżetu, bo stan jest ograniczony. Zdarzenia pod jego uczuciami nie są w stanie wymienione: stoją w focusie. Stan jest renderowany do kontekstu jako słowa i wagi, tak że na te same słowa tej samej osoby odpowiada inny autor, gdy istota niesie świeże zranienie, niż gdy nie niesie nic. Nie trzeba reguły w rodzaju „gdy zraniona, bądź chłodna”; przesunięcie robi swoje.

Focus to uwaga w sensie Kahnemana, ograniczony zasób: to, czego istota jest teraz świadoma, a nie historia rozmowy. Wszystko w nim ma postać emo-myśli, konstrukty też. Jedna linia trzyma wpisy w kolejności, w jakiej weszły w pole widzenia: uczucia, które istota przyniosła z wcześniej, emo-myśli z wymian, każde skojarzenie w miejscu, w którym zostało wywołane, i trzymane konstrukty według czasu, w którym przyszły na myśl. Widoczne są tylko cztery ostatnie wpisy; najstarszy wychodzi i jest zwijany do streszczenia. Wpis, którego uczucie jest w stanie, pokazuje tylko swoje zdarzenie. Człowiek nie trzyma w uwadze uczuć sprzed dziesięciu minut; te silne żyją dalej w stanie.

Streszczenie rozmowy jest częścią focusu i trzyma to, co opuściło uwagę. Zachowuje tylko epizody: co się stało, co powiedziano i co zostało przywołane, bez uczuć i bez wniosków. Jest prowadzone w dwóch częściach: co istota przyniosła sprzed tej rozmowy i

sama rozmowa. Zwijanie pozwala istocie przeżyć bardzo długie rozmowy, tak jak człowiek zachowuje fakty z rozmowy, a nie każde słowo i każde uczucie.

Stosowne dyspozycje to gotowość istoty na ten rodzaj sytuacji: wyuczone reguły, których sytuacja pasuje do tego, co dzieje się teraz, od zera do trzech. Są nieświadome: stoją poza focussem i działają tylko wtedy, gdy układana jest odpowiedź, to znaczy gdy istota działa.

3.5 Czyje słowa

Zapis niesie informację, kto go wytworzył. Autorem emo-myśli jest osoba, z imienia, albo sama istota; zapis dany przy narodzinach, gdy scenariusz takiego używa, jest oznaczony jako dany, a nie przeżyty. Epizod niesie to samo oznaczenie, a ponadto istnieje w dwóch postaciach. Gdy istota czyta wspomnienie, epizod jest pokazany opowiedziany, w trzeciej osobie i w czasie przeszłym: słowa drugiej osoby pozostają jej, a wcześniejsza odpowiedź istoty pozostaje przeszła. Tylko uczucie i sens, jaki z niego wynikł, są teraźniejszością istoty. Epizod taki, jak się zdarzył, służy do wyszukiwania i do bieżącej rozmowy, w której własna wymiana istoty sprzed chwili jest tym, co powiedziała. Tu i teraz wyznacza tę samą granicę przy bodźcu: to, co osoba mówi o istocie, jest jej twierdzeniem, ważnym wobec tego, co istota niesie. Bez tej granicy istocie czytającej swoje wspomnienia dosłownie można by powiedzieć „wczoraj mówiłaś, że go nienawidzisz”, a ona wzięłaby to zdanie za własne.

3.6 Procesy

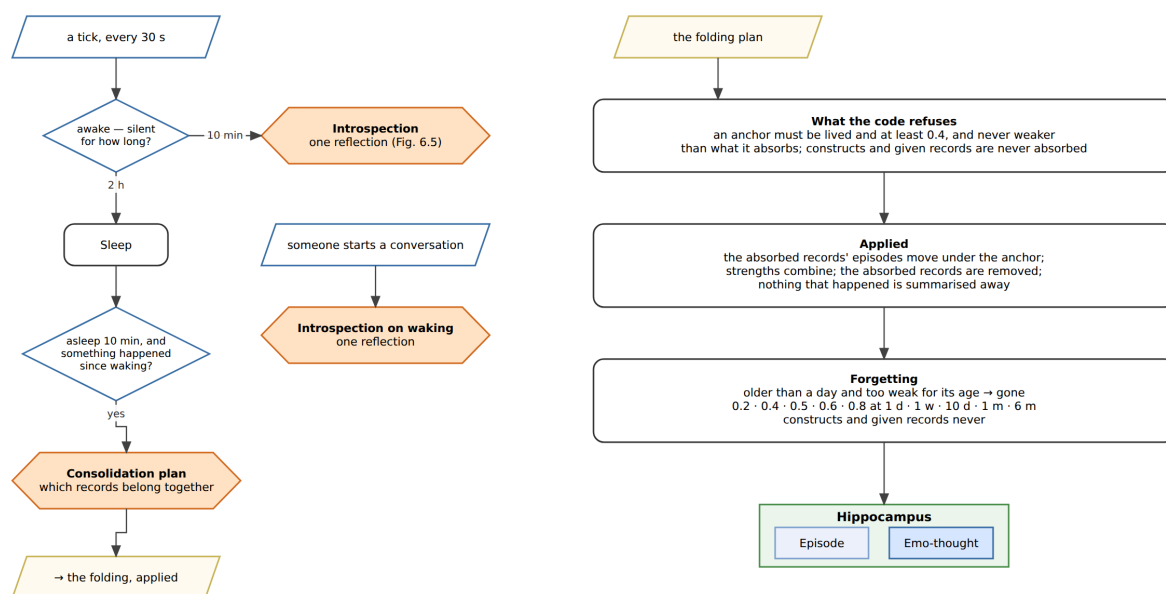
Tura to szybka ścieżka. Nadchodzi bodziec i dzieje się, co następuje, w tej kolejności. Wyuczone reguły pasujące do bodźca są wybierane do stosownych dyspozycji. Przywołanie wydobywa to, co bodziec przywiódł na myśl, do skojarzeń. Stan jest składany z najnowszych zapisów i najsilniejszych skojarzeń; uczucia, które wnosi z wcześniej, wchodzi do linii focusu. Generowanie odpowiedzi wytwarza działanie: odpowiedź, albo ciszę, albo wyszukiwanie, albo pauzę na namysł. Wyszukiwanie, albo jedno z własnych wyszukiwań silnika, gdy osoba jest nieznana lub pyta o jakąś chwilę, dodaje do skojarzeń i odpowiedź jest generowana ponownie. Pauza na namysł uruchamia najpierw refleksję (niżej). Ocena (appraisal) tworzy zapis, który zostanie zachowany, w drugim odczytaniu zdarzenia, ze wstrzymanym stanem, tak by zły nastrój zabarwiał odpowiedź, ale nie pamięć; to wybór tego, jaki rodzaj umysłu budować, a nie fakt o umysłach. Temperament nagina walencję i intensywność zapisu. Utrwalenie go zapisuje. W obrębie rozmowy chwila, której uczucie ma ten sam znak co przeżywany właśnie zapis, dołącza do niego jako kolejny epizod; zmiana znaku otwiera nowy zapis. Wspomnienie właśnie przywołane, które zgadza się znakiem z chwilą, odzyskuje siłę, nigdy ponad to, co jest odczuwane teraz; chwila zachowuje własny zapis. Zapis zachowuje uczucie, z którym powstał: jego walencja nigdy nie jest przepisywana, a zmienia się jego siła i to, co stoi obok niego. Zachowuje też datę tego, co się stało; przesuwają się tylko data, kiedy był ostatnio odczuwany. Wymiana dołącza do linii focusu, a po niej wspomnienia, które wywołała; gdy

linia trzyma więcej niż cztery wpisy, najstarsze opuszczają uwagę, a ich zdarzenia są zwiżane do streszczenia. Nic na tej ścieżce nie rozważa; stanowisko, jakie istota zajmuje wobec osoby, niosą emo-myśli, które o niej ma.

Refleksja to wolna ścieżka. Introspekcja uruchamia się przy czterech okazjach: gdy istota postanawia pomyśleć przed odpowiedzią, gdy osoba odchodzi, po ciszy i po przebudzeniu. Patrzy na to, co istota trzyma w umyśle, i na jej najnowsze zapisy, i tworzy jedną szczerą myśl w pierwszej osobie, bez żadnego nadanego kierunku. Myśl jest konstruktem poznawczym. Jej działanie może ciągnąć łańcuch dalej, przez kolejne pomyślenie albo przeszukanie pamięci, przez najwyżej trzy kroki, albo zakończyć go, tworząc konstrukt innego typu: przekonanie, decyzję, cel lub marzenie. To, co wytwarza wolna ścieżka, nie wraca do stanu odczuwanego. Jest trzymane według reguły slotu dostępności (§3.4) i stoi w focusie, gdzie czyta je szybka ścieżka. Samokierowanie, w tym modelu pojęciowym, jest tym, co obie ścieżki wytwarzają razem: konstrukt powstały na wolnej ścieżce jest trzymany w focusie, gdzie czyta go szybka ścieżka, tak że późniejszy akt ma adres w czymś, co istota napisała, a gdy zdarzenia, które go wytworzyły, wygasły, w niczym innym.

Odejście osoby jest osobnym bodźcem. Nie tworzy zapisu, uruchamia introspekcję i uruchamia uczenie dyspozycji: istota pyta, czego ta rozmowa nauczyła ją o tym, jak lepiej przetrwać takie spotkanie, nad tym samym kontekstem, z którego pracowała: swoim stanem i swoim focusiem ze streszczeniem. Zwykle niczego. Gdy czegoś, reguła jest zapisywana jako: gdy dzieje się coś takiego, zrób to, z uwzględnieniem znaku i siły najsilniejszego uczucia rozmowy, i umieszczana wobec reguł, które istota trzyma: ta sama reguła innymi słowami jest tak rozpoznawana po swojej sytuacji i swoim akcie, i silniejsza z dwóch zachowuje tekst; reguła sprzeczna osłabia starą i dodaje własną; reguła o czymś innym jest dodawana.

Sen przychodzi po długiej ciszy. Konsolidacja zwiż dzień: silne uczucia go kotwiczą, a epizody słabszych zapisów bliskich kotwicy w czasie lub w temacie przechodzą pod kotwicę, tak że nic, co się zdarzyło, nie zostaje streszczone do zniknięcia. Zapominanie usuwa to, co słabe jak na swój wiek, wraz z epizodami, poniżej progu, który rośnie z wiekiem; ze starych rzeczy zostają tylko silne emocje, a stare szczegóły tylko wtedy, gdy wiszą na jednej z nich. Konstrukty i zapisy dane przy narodzinach nigdy nie są wchłaniane ani zapominane. **Wygasanie** nie jest procesem snu, lecz czasu: siła zapisu spada z jego wiekiem ku dolnej granicy, nigdy do zera, uczucia negatywne wolniej, a spadek jest zapisywany do zapisu, gdy jest on następnym razem odczuwany, tak że nawrót ma co przywrócić. Same chwile i uczucie, z którym zapis powstał, nigdy nie są przepisywane. Mechanizm snu istnieje w silniku; życia opisane tutaj trwają dni i nie ćwiczą niczego poza zapominaniem. Opisujemy go, bo model pojęciowy jest bez niego niepełny, i niczego dla niego nie twierdzimy (Rys. 3.3).



Rys. 3.3. Między rozmowami: cisza prowadzi do refleksji, dłuższa cisza do snu, a sen do zwijania i zapominania. Model LLM proponuje, które zapisy należą do siebie; kod decyduje, co może służyć za kotwicę, a co nigdy nie może zostać wchłonięte ani zapomniane - ziarna i własne konstrukty owcy.

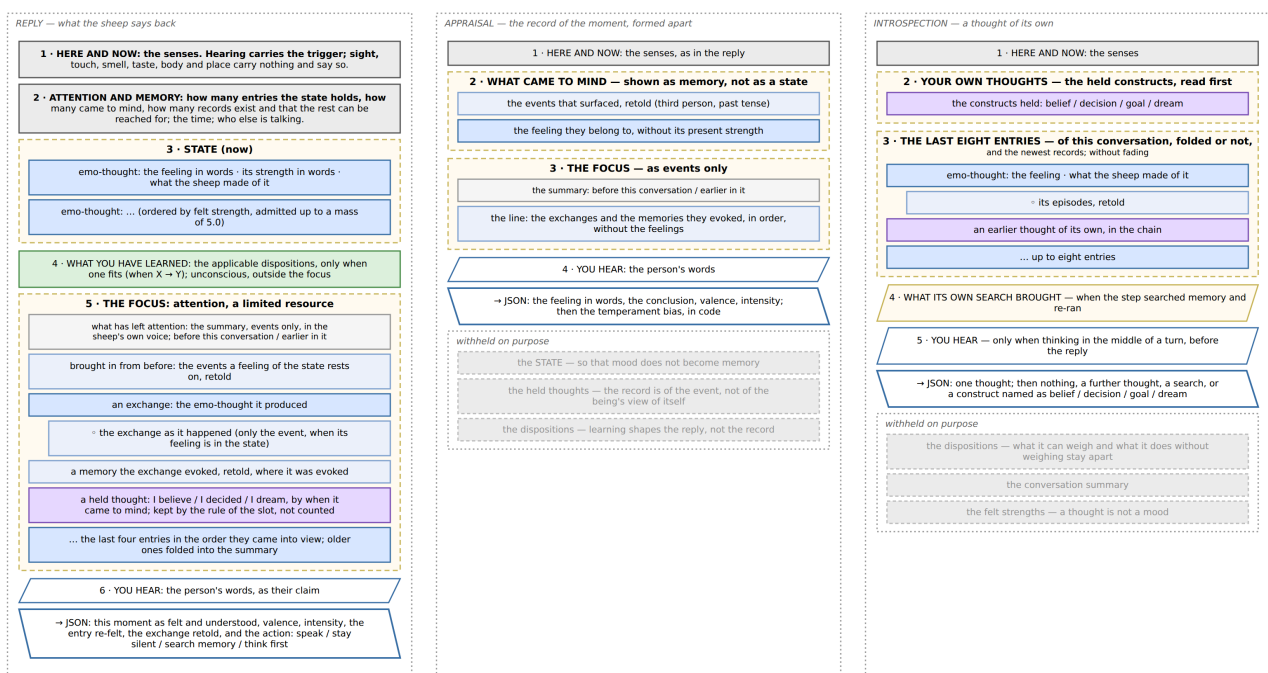
3.7 Konteksty: co widzi każde wywołanie

Kontekst to to, które kontenery widzi jedno wywołanie LLM. Rozdziały między kontekstami są projektem i każdy jest zamierzony: ocena jest tworzona bez stanu, tak by nastrój nie stawał się pamięcią; introspekcja nie widzi dyspozycji, tak by to, co istota może rozważyć, i to, co robi bez rozważania, pozostawały osobno; dyspozycje docierają do odpowiedzi i do niczego więcej, tak by uczenie zmieniało akt, a nie uczucie. Tabela podaje trzy wywołania istoty i wywołania pomocnicze.

wywołanie	widzi	wstrzymane
generowanie odpowiedzi	tu i teraz, zmysły; stan, uczucia z ich słowami i siłą; stosowne dyspozycje; focus: streszczenie tego, co opuściło uwagę, sprzed tej rozmowy i wcześniej w niej, potem linia czterech ostatnich wpisów w kolejności, w jakiej weszły w pole widzenia: uczucia przyniesione z wcześniej, opowiedziane; wymiany tak, jak się zdarzyły, każda pod swoim uczuciem, chyba że uczucie jest w stanie; wywołane wspomnienia, opowiedziane; trzymane konstrukty; bodziec	nic
ocena	tu i teraz; co przyszło na myśl w tej chwili, jako pamięć; streszczenie; linia focusu jako same zdarzenia; bodziec	stan, trzymane konstrukty, stosowne dyspozycje
introspekcja	tu i teraz; slot dostępności, na początku; ostatnie zapisy tej rozmowy, zwinięte lub nie, i hipokampu, z ich epizodami, bez wygasania; co przyniosło jej własne wyszukiwanie; bodziec, jeśli w środku tury; poprzednia myśl, w łańcuchu	stosowne dyspozycje, streszczenie, odczuwana siła
filtr skojarzeń	bodziec; kandydaci do przywołania, każdy przez zdarzenie, które go wydobyło	wszystko inne

wywołanie	widzi	wstrzymane
wybór dyspozycji	dyspozycje; dwie ostatnie wymiany osoby; bodziec; po jednej linii z trzech najnowszych wniosków lub własnych myśli, które istota niesie	stan
uczenie dyspozycji	kontekst, z którego istota pracowała, gdy osoba odeszła: stan; focus ze streszczeniem, opowiedziane epizody, na których opierają się uczucia, i jej trzymane myśli; to, co niesie; dyspozycje, do umieszczenia	nic
zwijanie (streszczenie)	jedna część streszczenia, sprzed tej rozmowy lub wcześniej w niej; zdarzenia wpisów opuszczających linię, od najstarszego	uczucia, wnioski, stan, trzymane konstrukty
konsolidacja	zapisy okresu czuwania z ich epizodami	konstrukty i zapisy dane, jako kandydaci do wchłonięcia

Rys. 3.4 rysuje to samo dla trzech wywołań, które liczą się najbardziej: co stoi w kontekście odpowiedzi, oceny i introspekcji, w kolejności, w jakiej czyta to model LLM, z zapisami zagnieżdżonymi w ich kontenerach i z oznaczeniem tego, co wstrzymane. Czytelnik, który chce odbudować warstwę, może to zrobić z tego rysunku i z §4.



Rys. 3.4. Trzy konteksty. Po lewej odpowiedź: zmysły, stan z jego emo-myślami, stosowne dyspozycje, focus: streszczenie tego, co opuściło uwagę (sprzed tej rozmowy, wcześniej w niej), potem jedna linia czterech ostatnich wpisów w kolejności, w jakiej weszły w pole widzenia (z czym istota przyszła, wymiany, wywołane wspomnienia, jej trzymane myśli), i to, co słyszy. Pośrodku ocena: te same zmysły i bodziec, to, co przyszło na myśl, pokazane jako pamięć, streszczenie i linia jako same zdarzenia; stan, trzymane myśli i dyspozycje wstrzymane. Po prawej introspekcja: najpierw slot, ostatnie zapisy z ich epizodami, co przyniosło jej własne wyszukiwanie, bodziec, jeśli w środku tury, i poprzednia myśl; dyspozycje, streszczenie i odczuwane siły wstrzymane.

3.8 Czym model pojęciowy nie jest

Model pojęciowy nie testuje teorii z §2 jako praw. To zbiór procesów wybranych dlatego, że idą za znanym wzorcem tego, jak doświadczenie staje się charakterem, zoperacjonalizowanych na zamrożonym modelu językowym i ocenianych przez to, co robi owca.

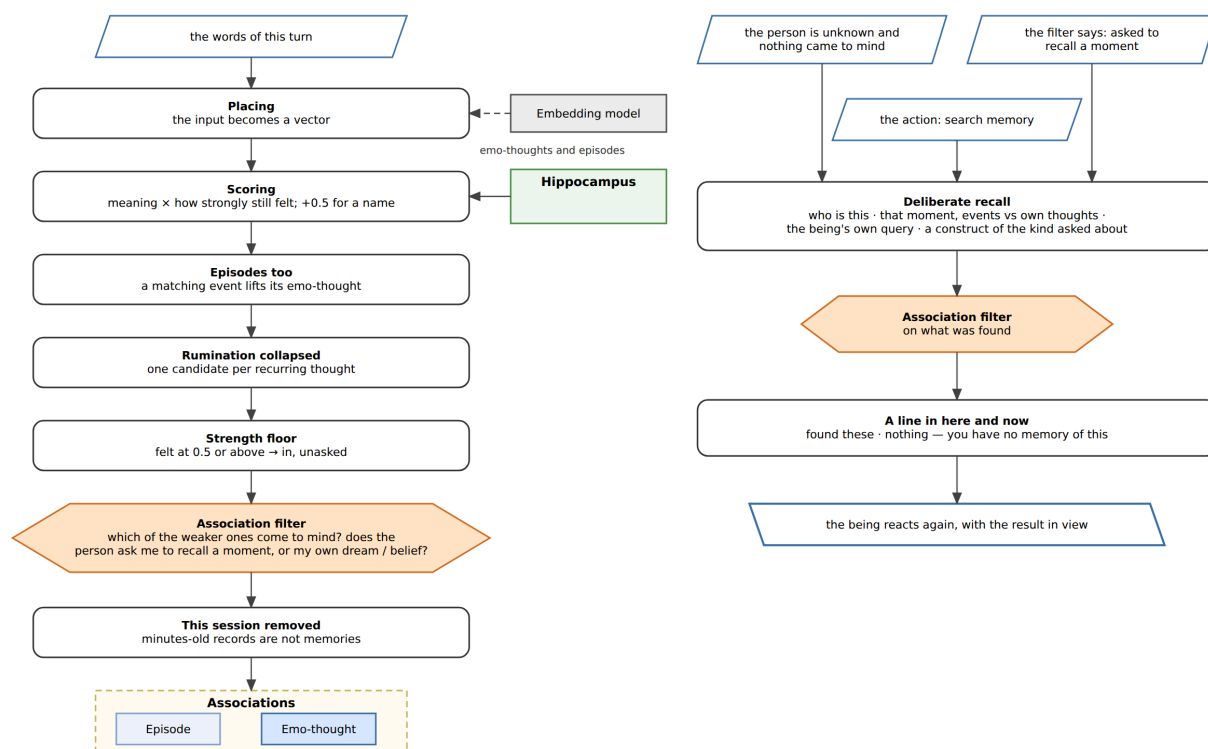
4. Architektura systemu i metoda

Ten opis podąża za kodem badanej wersji; daty poszczególnych zmian są w dzienniku zmian promptów i w dziennikach decyzji (Dodatek B). Każda liczba poniżej to wartość w działającej konfiguracji; sekcję zamyka tabela parametrów. Rysunki używają notacji z Rys. 3.1: flaga to osoba poza warstwą; pochyły prostokąt to bodziec lub wejście rysunku, prostokąt pochylony odwrotnie to działanie lub wyjście; szare pole to komponent, z którym warstwa rozmawia; kolorowy prostokąt to zapis, niebieski emo-mysł, bładoniebieski epizod, fioletowy konstrukt poznawczy, zielony dyspozycja; zielona ramka to kontener trwały, a bursztynowa przerywana ramka chwilowy; zaokrąglone pole to proces wykonywany przez kod, pomarańczowy sześciokąt proces wykonywany przez procesor w jednym wywołaniu LLM, romb decyzja w kodzie; szare przerywane pole oznacza miejsce, do którego coś nigdy nie dociera, a przerywana strzałka ścieżkę opcjonalną lub kolejny skok. Każdy rysunek jest na stronie diagramów (Dodatek I).

Rys. 4.1 mapuje domeny warstwy i to, co przechodzi między nimi w jednej turze: co jest spostrzegane, przywoływane, trzymane, składane w kontekst modelu LLM, mówione, odczuwane i zapisywane, oraz każde miejsce, w którym model LLM jest o coś pytany. Cztery z jego pól są otwarte na własnych rysunkach: przywołanie (Rys. 4.2), tworzenie zapisu (Rys. 4.4), introspekcja (Rys. 4.5) i dyspozycje (Rys. 4.6); tura jako pętla wywołań i skoków to Rys. 4.3. Dokładna sekwencja wywołań tury jest w śladzie kroków panelu badawczego (Dodatek D); tekst każdej operacji w postaci, w jakiej dociera do modelu LLM, jest w promptach (Dodatek A).

gubią nazwy własne. Wyszukiwanie obejmuje epizody, nie tylko emo-myśli: zdarzenie, którego fakt przypomina wejście, wprowadza swoje nadrzędne uczucie do kandydatów, z wynikiem równym podobieństwu zdarzenia razy odczuwana siła rodzica. Emocje bramkują i wzmacniają; samo wyszukiwanie biegnie po tym, co się zdarzyło.

Dziesięciu kandydatów z najwyższym wynikiem jest najpierw odduplikowanych, gdy reprezentują tę samą powracającą myśl (ruminacja to jedna natrętna myśl, nie dziewięć), a potem przekazanych do drugiego, małego wywołania LLM, filtru skojarzeń, który czyta wejście i kandydatów i wybiera tych, którzy naprawdę przyszliby człowiekowi na myśl; przeżywa najwyżej czterech. Jeden wyjątek chroni przed tym, by filtr przeważał nad siłą: zapisy odczuwane obecnie na 0.5 lub wyżej przechodzą bez pytania. Filtr jest mimo to zawsze pytany o resztę, bo czyta wejście jeszcze pod jednym kątem: czy osoba prosi owcę o przywołanie chwili z jej własnej przeszłości (§4.5). Zapisy napisane minuty wcześniej w tej samej rozmowie są wyłączone z tego, co liczy się jako pamięć: są już widoczne jako rozmowa, a podane jako wspomnienia czytają się jak wspólna przeszłość z nieznanym (Rys. 4.2).



Rys. 4.2. Przywołanie: znaczenie pomnożone przez obecną siłę, imiona podbite, epizody przeszukiwane obok emocji, ruminacja zwinięta do jednego reprezentanta, wszystko odczuwane jeszcze na 0.5 wpuszczone bez pytania, reszta przekazana filtrowi skojarzeń. Poniżej trzy celowe wyszukiwania, każde kończące się drugim wywołaniem filtru i linią w zmysłach - łącznie z wypowiedzianym null.

4.3 Kontenery: stan, focus i streszczenie rozmowy

Stan na turę czerpie z dwóch źródeł: pięciu najnowszych zapisów z życia owcy, niezależnie od tego, kogo dotyczą, w ich obecnej sile i bez względu na wiek, oraz skojarzeń, które wejście przywiodło na myśl. Ostatnia rzecz, która przydarzyła się owcy,

jest w jej teraźniejszości, dopóki nie wyprze jej coś nowszego; po tygodniu samotności jest wygaszona, i wciąż tam jest. Zapisy są porządkowane według bieżącej odczuwanej siły i wpuszczane, aż ich łączna masa osiągnie 5.0; umysł nie trzyma nieograniczonej ilości afektu naraz. Każde wpuszczone uczucie pojawia się ze swoimi słowami i swoją siłą. Epizody, na których się opiera, nie są pod nim wymienione: stoją w focusie. Stan to to, co owca czuje; focus to to, czego jest świadoma. Focus ma dwie części, renderowane w tej kolejności po stanie: streszczenie tego, co opuściło uwagę, pod własnym nagłówkiem, w dwóch częściach („sprzed tej rozmowy” i „wcześniej w tej rozmowie”); potem linia. Linia trzyma cztery ostatnie wpisy w kolejności, w jakiej weszły w pole widzenia: uczucie przyniesione z wcześniej wchodzi do linii, opowiedziane, gdy po raz pierwszy wchodzi do stanu, a raz zwinięte nie wraca ze stanem; każda wymiana tej rozmowy stoi tak, jak się zdarzyła, pod uczuciem, które wytworzyła, chyba że to uczucie jest w stanie, a wtedy pokazana jest tylko wymiana; każde wspomnienie wywołane przez wymianę stoi po niej, opowiedziane, a wspomnienie wywołane ponownie po tym, jak wyszło z pola widzenia, wraca do linii; konstrukty trzymane przez slot stoją według czasu, w którym powstały lub zostały przywołane na myśl, po jednej linii każdy, ze swoim rodzajem, i nie liczą się do czterech. Zamrożone odtworzenie dwudziestu kontekstów pokazało, że na uczucie ze zdarzeniami pod nim odpowiada się jak na uczucie, a na to samo uczucie ze zdarzeniami w focusie jak na sytuację (poprzedni dzień był niesiony w dziesięciu z dziesięciu kontekstów wobec trzech z dziesięciu).

Gdy wymiana jest zachowana, a wspomnienia, które wywołała, weszły, wpisy ponad cztery opuszczają uwagę. Jedno wywołanie na część streszczenia przepisuje tę część w całości: zachowuje to, co część trzymała, i dodaje zdarzenia z wychodzących wpisów, w pierwszej osobie, tylko wymienione zdarzenia, z wywołanym wspomnieniem jako „przywołałam ...”, bez niczego dodanego, bez uczuć i wniosków, najwyżej osiem zdań. Jeśli wywołanie się nie powiedzie, wpisy zostają w polu widzenia i zwijają się w następnej turze. Stan niesie emocję.

4.4 Co widzi model LLM: interfejs

Kontekst modelu LLM nie zawiera żadnej instrukcji, co czuć ani jak działać. Nazywa tylko to, co model LLM czyta i co ma zwrócić. Tekst systemowy, poza skalą intensywności, która po nim następuje, brzmi:

TWÓJ STAN poniżej to to, co czujesz teraz: każde uczucie z myślą, z którą przyszło, i sensem, jaki mu nadałaś, oraz wspomnienia, które przyszły na myśl. Niektóre wpisy to twoje własne wcześniejsze myśli, przeświadczenia, postanowienia, chęci i pragnienia. Odpowiadasz z niego.

Twoja pamięć zawiera to, co ci się przydarzyło, i nic więcej. Czego w niej nie ma, nie przydarzyło ci się.

Zwracasz dwie rzeczy: tę chwilę, tak jak ją poczułaś i zrozumiałaś, i to, co odpowiadasz.

Każdą wartość pisz w języku słów, które słyszysz. Nazwy pól zostają po angielsku. Zwróć tylko JSON.

Dalej następuje sześciolinijkowa skala, która definiuje intensywność do zwrócenia: 0.0 ślad, 0.2 słaba, 0.4 umiarkowana, 0.6 silna, 0.8 bardzo silna, 1.0 przytłaczająca. Tekst użytkownika składa się z tury w tej kolejności:

1. **Tu i teraz: zmysły.** Słuch niesie bodziec: słowa rozmówcy albo nic, jeśli ktoś tylko wszedł lub wyszedł. Wzrok, dotyk, węch, smak, ciało i miejsce nie niosą nic i mówią o tym: „nie widzisz nic prócz czerni, jakbyś była ślepa” („you see nothing but black, like being blind”).
2. **Tu i teraz: uwaga i pamięć.** Ile wpisów ma stan, ile wspomnień przyszło na myśl, ile zapisów ma pamięć, że to, co pod stanem, to to, co właśnie teraz wróciło na myśl, a reszta jest i można po nią sięgnąć wyszukiwaniem; potem czas i kto jeszcze rozmawia z owcą równolegle.
3. **Stan.** Każde uczucie z jego siłą wyrażoną słowami, jak było odczuwane i co owca z niego zrozumiała.
4. **Pasujące dyspozycje.** Od zera do trzech wyuczonych reguł, jako reguły, pod własnym nagłówkiem, tylko gdy któraś pasuje do sytuacji (§4.9); działają na odpowiedź i nie są częścią focusu.
5. **Focus.** Najpierw streszczenie tego, co opuściło uwagę, głosem samej owcy, w dwóch częściach: przed tą rozmową i wcześniej w niej; potem linia ostatnich czterech wpisów w kolejności, w jakiej pojawiły się w polu widzenia: uczucia, z którymi owca weszła, opowiedziane; wymiany, tak jak się wydarzyły, każda pod swoim uczuciem, chyba że uczucie jest w stanie; wspomnienia, które każda wymiana wywołała, opowiedziane; jej własne trzymane myśli, każda ze swoim rodzajem (wierzę, zdecydowałam, marzę).
6. **Słowa rozmówcy,** jako to, co niesie słuch: cokolwiek osoba mówi o owcy, o sobie lub o tym, co tu jest, to jej twierdzenie, do zważenia wobec tego, co owca nosi w sobie.
7. **Kształt JSON.** Uczucie tej chwili, interpretacja owcy, walencja, intensywność, opowiedzenie wymiany i działanie: mówić, z odpowiedzią, albo milczeć, albo przeszukać pamięć, albo najpierw pomyśleć.

Trzy rzeczy trzeba rozdzielić, gdy ten interfejs nazywa się wolnym od instrukcji. Owca nie dostaje biografii ani osobowości; nic nie mówi jej, kim jest. Nie dostaje reakcji; nic nie mówi jej, co czuć ani co mówić. Dostaje ramę: kontekst mówi o tym, co ty czujesz i co ty nosisz w sobie, a refleksja może zakończyć się przekonaniem, decyzją, celem lub marzeniem. Rama umożliwia „ja”; nie dostarcza go. W tym samym sensie „niepytana” znaczy bez prośby ze strony osoby: cicha refleksja biegnie według harmonogramu silnika, nie według pytania gościa. Co widzą pozostałe wywołania i co jest przed każdym z nich ukryte, pokazuje tabela w §3.7. Zdanie tekstu systemowego o własnych wpisach owcy jest bezczynne: te wpisy stoją w focusie, nie w stanie.

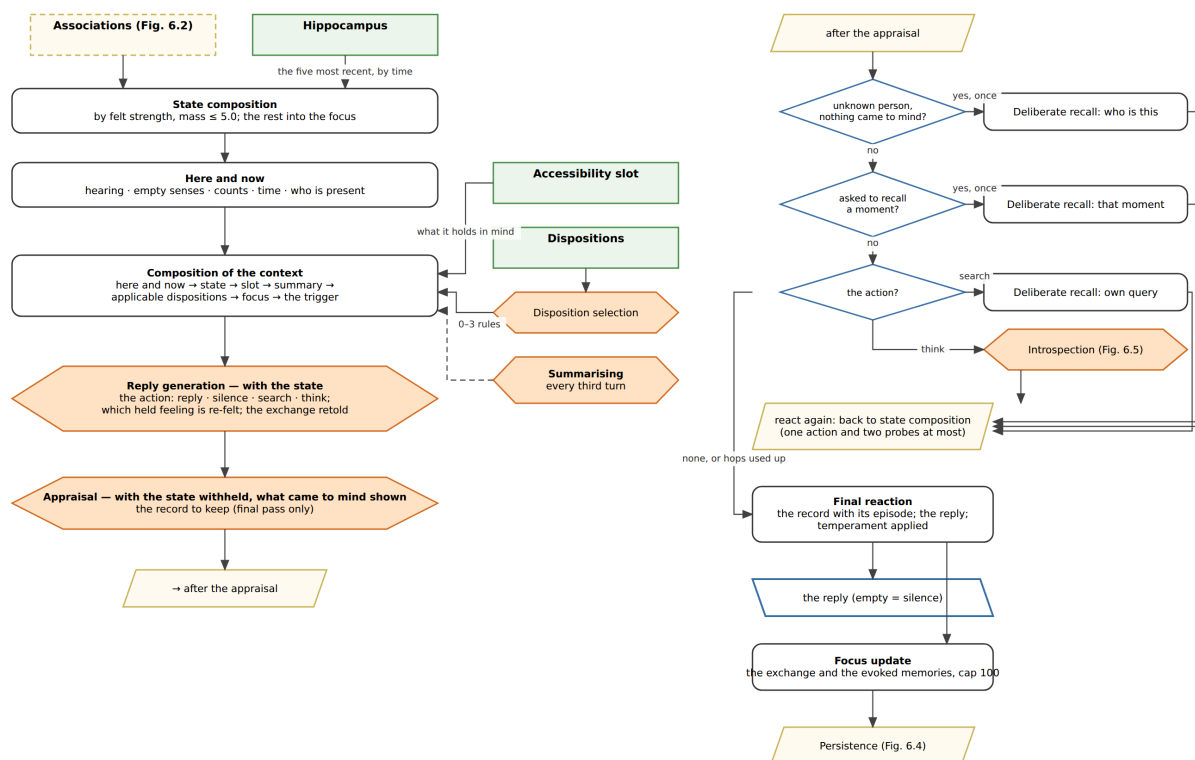
Ten interfejs jest wynikiem przycinania. Jego poprzednik zawierał tysiąc dziewięćset słów reguł: że owca „nie jest uprzejmym, domyślnie włączonym asystentem” („not a polite default-on assistant”), że powitanie nie może zresetować jej nastroju, że wspomnienia o obecnej osobie są szczególnie wyraziste, że postanowienie „NADPISUJE twój domyślny

impuls" („OVERRIDES your default impulse"), a decyzja o wycofaniu się znaczy „bądź oschła, zimna, ostrożna, zwięzła albo milcz" („be curt, cold, guarded, brief, or silent"). Każdą dodano w odpowiedzi na zaobserwowaną porażkę wcześniejszego modelu LLM lub wersji silnika. Zgodnie z zasadą, że każde zdanie widoczne dla LLM musi zasłużyć na swoje miejsce (reguły interfejsu projektu, Dodatek B), każdą usunięto, chyba że zamrożone odtworzenie pokazało, że jest konieczna; żadna nie była. Pełny tekst każdego typu wywołania i dziennik każdej zmiany z jej dowodami są w Dodatku.

4.5 Tura

Jedna tura to jedno wejście. Po przywołaniu i zbudowaniu stanu główne wywołanie wytwarza działanie: wypowiedzianą odpowiedź albo milczenie, albo przeszukanie pamięci z własnym zapytaniem owcy, albo pauzę na namysł przed odpowiedzią; wraz z nim tymczasową emo-myśl. Drugie wywołanie o tym samym zdarzeniu, z celowo ukrytym stanem i z tym, co przyszło na myśl, pokazanym mu jako pamięć, formuje zapis, który zostanie zachowany; ta izolacja powstrzymuje zły nastrój przed zamienianiem wszystkiego, co usłyszane, w kolejne złe wspomnienie, a zarazem żaden zapis nie powstaje w próżni; jest to wybór projektowy, a nie fakt o umysłach. Oba wywołania dostają skalę intensywności; pierwsze wytwarza opowiadanie.

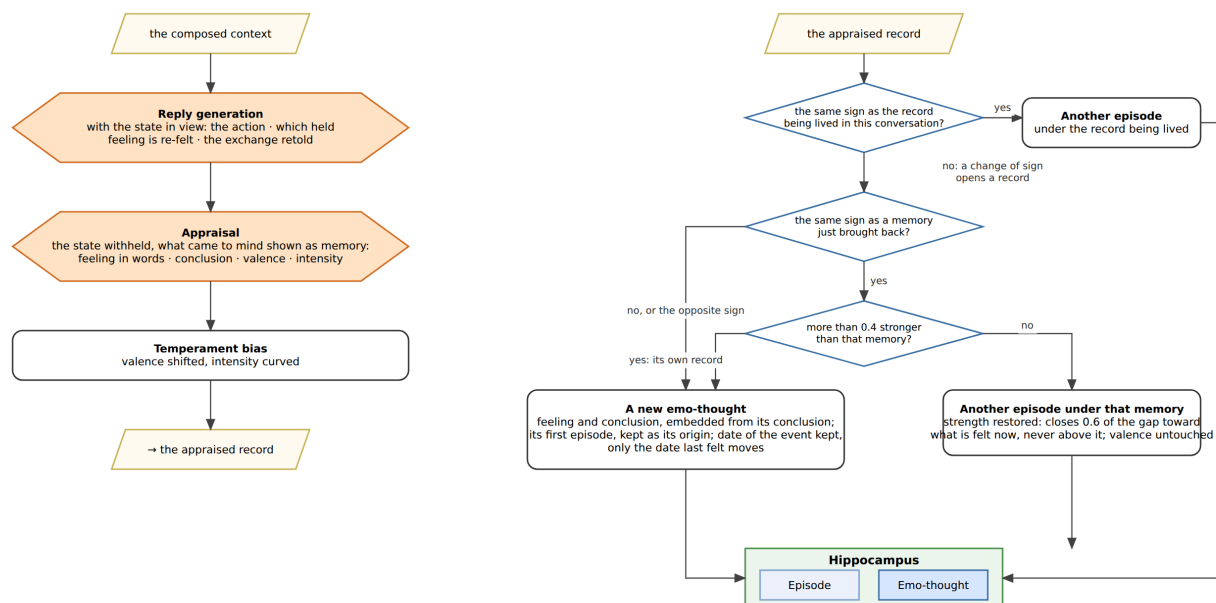
Dwie sondy biegną jako własne akty owcy, bez instrukcji. Jeśli odzywa się nieznana osoba i nic nie przychodzi na myśl, silnik wyszukuje tę osobę i ponownie uruchamia reakcję z wypowiedzianym wynikiem w zmyśle pamięci: „przeszukałaś pamięć w poszukiwaniu X; nic nie znaleziono; nie masz o tym żadnego wspomnienia" („you searched your memory for X; nothing found; you have no memory of this"). Jeśli filtr skojarzeń zgłasza prośbę o przywołanie chwili, miejsca lub sceny z przeszłości owcy, silnik jej szuka, oddziela zdarzenia od własnych późniejszych myśli owcy i ponownie uruchamia reakcję z wynikiem: znalezione zdarzenia, nic, albo „nie ma takiej chwili; pojawiły się tylko twoje własne późniejsze myśli na ten temat" („no such moment; only your own later thoughts about the theme came up"). Ta druga sonda ugruntowuje część odpowiedzi na ciepłą prośbę o scenę, ale nie resztę; zmyślona reszta jest opisana w §9. Jeśli wywołanie odpowiedzi, nad kontekstem tury, zwróciło działanie „najpierw pomyśleć", przed odpowiedzią biegnie introspekcja (§4.7). Tura pozwala na jeden skok działania, wyszukiwanie albo pauzę, i do dwóch skoków sond; zapis formuje się tylko w ostatnim przejściu i ta reakcja zostaje zachowana (Rys. 4.3).



Rys. 4.3. Tura jako pętla wywołań i skoków: budowanie stanu, blok tu-i-teraz, wywołanie odpowiedzi, ocena odizolowana, dwie sondy, skoki wyszukiwania i namysłu, które ponownie uruchamiają reakcję, oraz końcowa reakcja, która zostaje zachowana.

4.6 Trwałość i rekonsolidacja

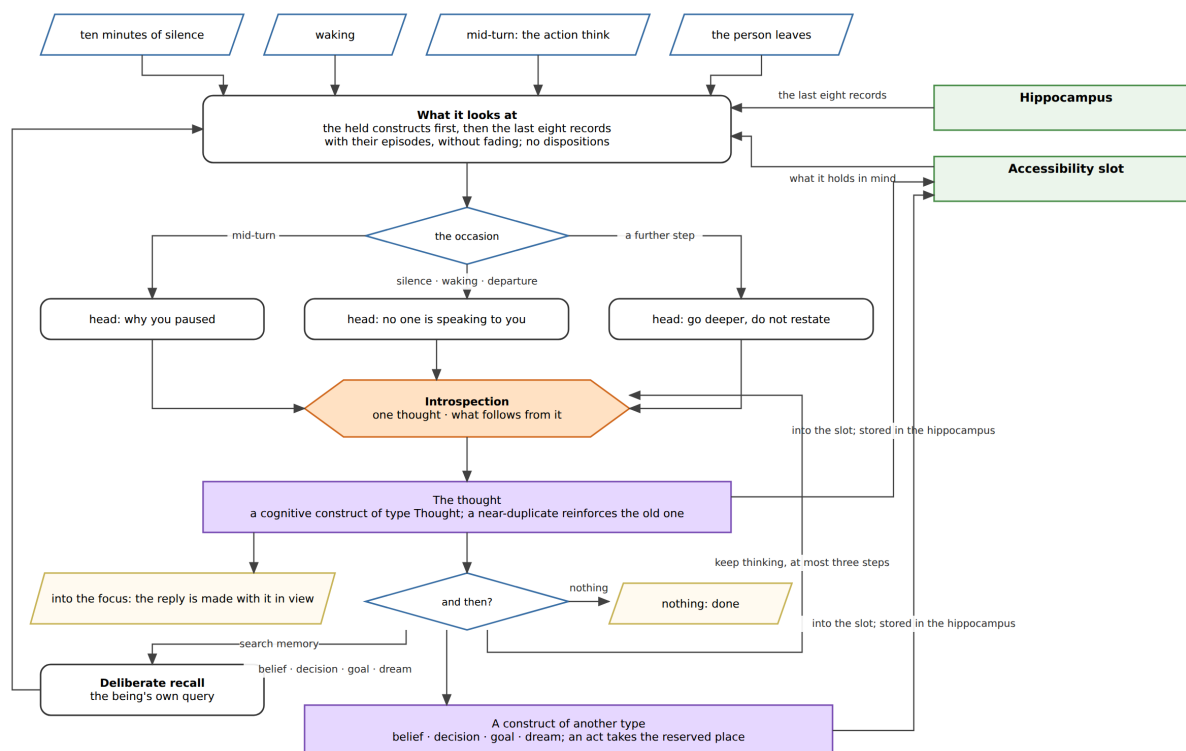
Jeśli główne wywołanie rozpozna nowe uczucie, zapis dostaje osadzenie i jest zapisywany ze swoim pierwszym epizodem, a zapis pamięta ten epizod jako swoje źródło. W obrębie rozmowy kolejna chwila, której uczucie ma ten sam znak, dołącza do przeżywanego zapisu jako kolejny epizod; zmiana znaku otwiera nowy zapis, tak że rozmowa, w której ktoś rani, a potem przeprasza, zostawia dwa zapisy, a nie jeden przesunięty do połowy. W przeciwnym razie chwila formuje własny zapis, a najsilniejsze przeżyte wspomnienie tego samego znaku, które ta tura przywiodła na myśl, wraca: jego siła zamyka sześć dziesiątych odstępu do tego, co jest odczuwane teraz, i nigdy nie wznosi się ponad to, tak że wspomnienie straty nie boli bardziej, niż bolała strata. Wspomnienie odzyskuje siłę, ale nie przejmuje nowego zdarzenia: wyśmianie za stratę nie jest nawrotem żałoby, a złożony pod starym wspomnieniem nowy cios byłby czytany jako stary smutek. Wspomnienie przeciwnego znaku nigdy tą drogą nie wraca. Walencja zapisu nigdy nie jest przepisywana: pocieszenie nie zamienia straty w przyjemność, stoi obok niej jako własny zapis. Siła opada między chwilami, w których zapis jest odczuwany (§4.2), a spadek jest zapisywany w zapisie, gdy jest on odczuwany następnym razem. Data tego, co się wydarzyło, pozostaje, jak była; przesuwa się tylko data, kiedy zapis był ostatnio odczuwany. Pogłębianie się powtarzanej krzywdy, uzdrowienie przez pocieszenie i utrata więzi po zdradzie wynikają z tego, co stoi obok zapisu, i z tego, jak silnie każde z nich jest wciąż odczuwane, bez osobnych przypadków (Rys. 4.4).



Rys. 4.4. Jak wymiana staje się zapisem: wywołanie odpowiedzi pisze słowa i opowiadanie; drugie wywołanie z ukrytym stanem pisze, co było odczuwane; temperament to zakrzywia; potem powstaje nowy zapis albo wymiana dołącza do zapisu przeżywanego w tej rozmowie, a wspomnienie, które chwila przywiodła, odzyskuje siłę. Same chwile i uczucie, z którym zapis powstał, nigdy nie są przepisywane.

4.7 Introspekcja

Introspekcja to osobne wywołanie, którego tekst nazywa operację i nic więcej: to chwila myślenia, nie odpowiedź; stan poniżej to to, co owca czuje teraz; napisz jedną własną myśl o tym, co się wyróżnia; nie zmyślaj zdarzeń, których nie ma w twojej pamięci; jeśli wynika z tego przeświadczenie, postanowienie, chęć lub pragnienie, nazwij je jako przekonanie, decyzję, cel lub marzenie, w przeciwnym razie nic. Jej kontekst to osiem wpisów: ostatnie wpisy tej rozmowy, łącznie z tymi już zwiniętymi poza focus, dopełnione najnowszymi zapisami. Biegnie, gdy wywołanie odpowiedzi zwróci działanie „najpierw pomyśleć”, w środku rozmowy, sama z siebie po zakończeniu sesji (od jednej do trzech myśli w ciszy, opcja scenariusza) i po dziesięciu minutach ciszy. Krok może skończyć się niczym, kolejną myślą, przeszukaniem pamięci, które uruchamia go ponownie, albo konstruktem. Myśli i pozostałe konstrukty są przechowywane jako własne zapisy owcy, deduplikowane semantycznie, i nie są noszone jak uczucia: przy włączonym słocie (§4.10) nic, co owca myśli o sobie, nie wchodzi do odczuwanego stanu ani do skojarzeń. Trzyma je reguła slotu: odpowiedź czyta je w focusie, następna introspekcja na początku swojego kontekstu (Rys. 4.5).



Rys. 4.5. Introspekcja: ta sama operacja w trzech okolicznościach - pauza w środku rozmowy, cisza, przebudzenie - nad ośmioma wpisami, kończąca się niczym, kolejnym krokiem, wyszukiwaniem albo konstruktem umieszczonym w slocie i w magazynie.

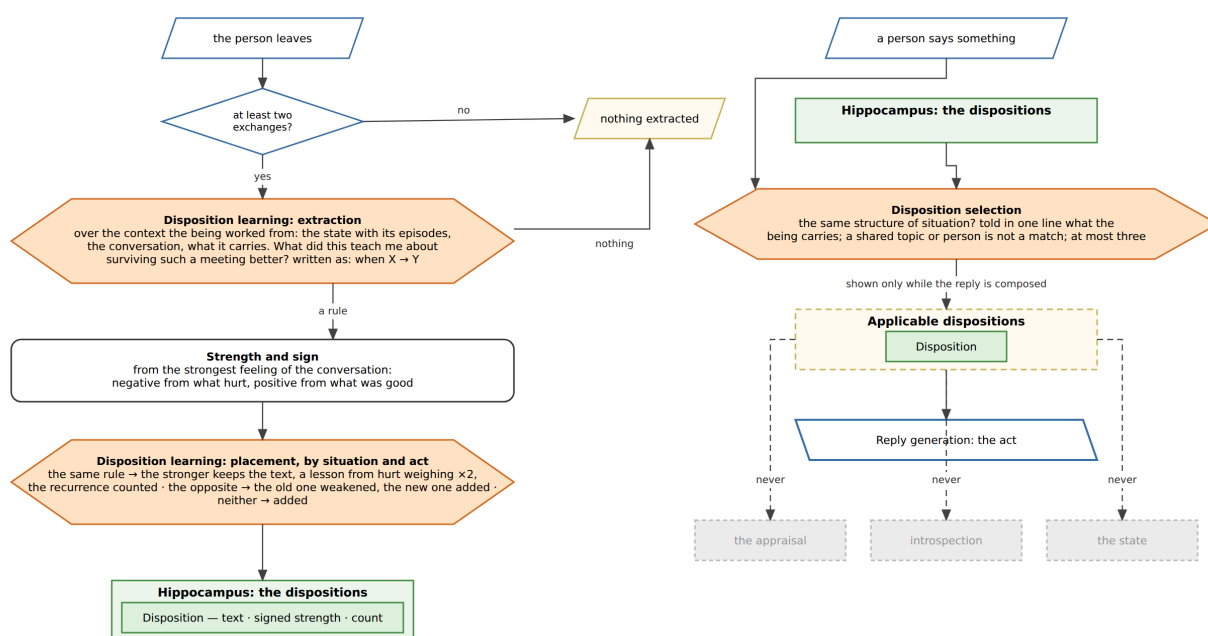
4.8 Sen

Sen to osobna operacja i nie wytwarza reguł. Po dwóch godzinach bezczynności silnik zwija zapisy leżące blisko w czasie silnego uczucia w to uczucie: model LLM proponuje grupy, a kod przenosi epizody wchłoniętych zapisów pod kotwicę i łączy ich intensywności, tak że nic, co się wydarzyło, nie znika w streszczeniu. Potem usuwa to, co jest za słabe jak na swój wiek, wraz z epizodami. Zegar scenariusza przeskakuje o dzień między sesjami i uruchamia zapominanie, ale własny harmonogram silnika liczy rzeczywisty upływ czasu, więc w opisanych życiach zwijanie nie biegło; biegło tylko zapominanie.

4.9 Wyuczone dyspozycje

Dyspozycja to to, czego owca nauczyła się o spotykaniu pewnego rodzaju sytuacji, by lepiej ją przetrwać. Jest wydobywana raz, gdy rozmowa się zamyka, tylko po co najmniej dwóch wymianach, z kontekstu, z którego owca pracowała w tej rozmowie: jej stanu, jej focusu ze streszczeniem i tego, co nosi w sobie. Jedno wywołanie pyta, czego ta rozmowa nauczyła owcę o lepszym przetrwaniu takiego spotkania, zapisane jako jedna reguła postaci *gdy dzieje się coś takiego* → *to właśnie robi owca*, a zwykła odpowiedź brzmi: nic. Reguła dostaje siłę ze znakiem od najsilniejszego uczucia rozmowy, która ją wytworzyła, ujemną, gdy lekcja wzięła się z tego, co bolało. Model LLM nie jest pytany, które to było: poproszony o nazwanie go, w ogóle odmawia napisania lekcji z krzywdy (§6.5); walencja,

którą silnik już zapisał, sama jest sygnałem. Drugie wywołanie umieszcza regułę wobec bazy według jej sytuacji i jej aktu. Ta sama reguła innymi słowami jest uznawana za tę samą i silniejsza z dwóch zachowuje tekst, przy czym lekcja z krzywdy waży dwa razy tyle co lekcja tej samej wielkości z dobra; reguła, która mówi coś przeciwnego dla tego samego rodzaju sytuacji, osłabia starą i jest dodawana; reguła o czymś innym jest dodawana. Gdy przychodzi wejście, a baza nie jest pusta, jedno wywołanie wybiera dyspozycje, których sytuacja pasuje do struktury obecnej, przy czym żadna to norma, a trzy to najwięcej, i dostaje w jednej linii informację, co owca nosi w sobie, tak by pytanie, które dotyka rany, nie było czytane jako niewinne. Tylko wywołanie odpowiedzi widzi wybrane linie, pod własnym nagłówkiem po stanie; ani wywołanie formujące zapis, ani introspekcja ich nie widzą i nigdy nie wchodzi one do stanu. W zamrożonym odtworzeniu umieszczenie tej samej linii w stanie zmieniło zapisane uczucie wobec przeprosin z -0.30 na -0.60 i zapisałoby surowsze wspomnienie; umieszczona tylko w budowaniu odpowiedzi, zostawiła uczucie bez zmian i zmieniła odpowiedź (Rys. 4.6).



Rys. 4.6. Wyuczone dyspozycje: bramka dwóch wymian, wydobywanie i umieszczenie, gdy rozmowa się zamyka; wybór na bieżącym wejściu; jedno miejsce, w którym wynik jest pokazywany, i trzy miejsca, do których nigdy nie dociera - ocena, introspekcja i stan.

4.10 Slot dostępności i przywołanie po rodzaju

Slot trzyma trzy własne zapisy owcy (myśli z introspekcji, przekonania, decyzje, cele lub marzenia) jako osobną tabelę odniesień do własnych zapisów owcy, tak że przetrwa koniec rozmowy. Zapis wchodzi, gdy powstaje lub zostaje przywołany; wychodzi, gdy zostaje wyparty albo po siedmiu dniach bez jednego i drugiego. Jedno miejsce ma pierwszeństwo dla rodzajów, według których owca działa lub po które sięga, decyzji, celów i marzeń: nowszy zastępuje trzymany, a przychodzący akt zawsze znajduje swoje miejsce, zabierając je najstarszemu przekonaniu, jeśli przekonania zapełniły wszystkie trzy, gdy żaden akt nie był trzymany. Przekonania i myśli wchodzi na pozostałe miejsca

tylko wtedy, gdy są co najmniej tak silne jak najsłabsze tam trzymane, i wtedy wychodzi najstarsze. Regułę wybrano, symulując pięć alternatyw na historiach konstruktów dwudziestu ukończonych żyć: w każdym wspólnym pojemniku decyzja, zapisywana na 0.65–0.7, podczas gdy przekonania są zapisywane na 0.85–1.0, była wypierana przez następne przekonanie i na koniec żadnego życia nie była trzymana; z zarezerwowanym miejscem była trzymana we wszystkich ośmiu życiach, które jakąś miały (Dodatek F). Introspekcja widzi trzymane zapisy najpierw, po jednym w linii, każdy ze swoim rodzajem („Wierzę:", „Zdecydowałam:", „Marzę:"; w oryginale „I believe:", „I decided:", „I dream:"); odpowiedź widzi każdy z nich w linii focusu, według czasu, gdy powstał lub wrócił na myśl, ze swoim rodzajem; w testach zamrożonych z §6.4 stały w osobnej sekcji nad focussem, YOUR OWN THOUGHTS; odizolowane formowanie zapisu nie widzi żadnego.

Przywołanie po rodzaju uzupełnia przywołanie semantyczne: gdy filtr skojarzeń rozpozna pytanie o to, o czym owca marzy, w co wierzy, co zdecydowała lub czego chce, silnik wykonuje jedno celowe wyszukiwanie wśród jej własnych zapisów tego rodzaju, oddane jako linia w zmysłach, jak pozostałe sondy. Przywołanie semantyczne nie potrafiło tego zrobić niezawodnie: na magazynie produkcyjnym marzenie wpływało tylko dla angielskiej formy pytania, nigdy dla polskiej.

4.11 Wywołania i parametry

Zwykła tura używa trzech wywołań czatu (filtr, reakcja, odizolowany zapis) plus, gdy linia focusu jest już pełna, wywołania zwijającego, gdy wpisy ją opuszczają (około 0.6 na turę w życiu „Rana"); tura z sondą lub introspekcją używa od pięciu do dziesięciu. Gdy osoba wychodzi, jedno wywołanie pyta, czy rozmowa nauczyła owcę czegoś, co da się uogólnić (§4.9); gdy baza dyspozycji nie jest pusta, jedno wywołanie na turę wybiera te, które pasują. Zmierzone na dwóch życiach z §5 z zapisanym każdym wywołaniem (Dodatek F): tura odpowiedzi wykonywała medianę sześciu wywołań (od czterech do dziewięciu), czytała medianę 28–31 tysięcy znaków kontekstu w tych wywołaniach (z grubsza 7–8 tysięcy tokenów, największe pojedyncze wywołanie około 13 tysięcy znaków), zwracała około 4 tysięcy znaków i zajmowała 21–26 sekund czasu rzeczywistego przy wywołaniach uruchamianych po kolei wobec hostowanego modelu LLM; trzydniowe życie z około trzydziestu tur wykonało 91 i 111 wywołań i przeczytało łącznie około 540 tysięcy znaków. Warstwa jest cienka w swoim interfejsie i kodzie, nie w koszcie na turę, który jest kilka razy większy niż koszt chatbota odpowiadającego z transkryptu.

Owca i każdy skryptowany rozmówca używają claude-haiku-4-5; osadzenia to nomic-embed-text v1.5, lokalnie.

parametr	wartość	gdzie
ostatnie zapisy w stanie	5, z obecną siłą, niezależnie od wieku	4.3
kandydaci do przywołania / zachowani przez filtr	10 / ≤ 4	4.2
premia za imię	+0.5	4.2

parametr	wartość	gdzie
wygasanie	dno 0.3, $\tau = 3$ dni, ujemne $\times 1.5$ wolniej	4.2
dno siły (filtr nie może zawetować)	0.5	4.2
przywołanie po epizodach	włączone	4.2
budżet masy stanu	5.0	4.3
linia focusu	4 wpisy; trzymane konstrukty nieliczone; starsze wpisy zwinięte do streszczenia	4.3
streszczenie	dwie części, przed tą rozmową i w niej; każda przepisywana w całości, gdy wpisy opuszczają linię; najwyżej 8 zdań	4.3
odizolowane formowanie zapisu	włączone	4.5
sondy (nieznana osoba, prośba o scenę)	włączone	4.5
dołączanie w obrębie rozmowy	ten sam znak dołącza jako epizod; zmiana znaku otwiera zapis	4.6
nawrót przywraca siłę o	0.6 odstępu do tego, co jest odczuwane teraz, nigdy powyżej	4.6
walencja zapisu	nigdy nie przepisywana	4.6
przyjścia i wyjścia	nie tworzą zapisu	3.3
epizody opowiedziane (trzecia osoba, czas przeszły)	włączone; z wywołania odpowiedzi	4.1
epizody w kontekście odpowiedzi	w linii focusu tam, gdzie pojawiły się w polu widzenia, opowiedziane, nie pod uczuciami	4.3
wyuczone dyspozycje	włączone; wydobyte przy zamknięciu rozmowy z kontekstu, z którego owca pracowała, jako <i>gdy</i> $X \rightarrow Y$; siła ze znakiem od najsilniejszego uczucia rozmowy; lekcja z krzywdy waży $\times 2.0$ przy kolizji; silniejsza zachowuje tekst; wybór ≤ 3 na turę, pokazywany tylko wywołaniu odpowiedzi	4.9
kontekst introspekcji / dodatkowe skoki na turę	8 wpisów / 1 działanie + 2 sondy	4.5, 4.7
slot dostępności	włączony; 3 miejsca, 1 z pierwszeństwem dla decyzji/celu/marzenia, 7-dniowy zanik, dynamiczny próg na wolnych miejscach	4.10
przywołanie po rodzaju	włączone; jedno wyszukiwanie wśród własnych zapisów pytanego rodzaju	4.10
sen (zwijanie i zapominanie) po / beczynna introspekcja po	2 h / 10 min	4.8, 4.7
temperament każdej mierzonej owcy	przesunięcie walencji 0.0, wzmocnienie intensywności 1.0	3.2

parametr	wartość	gdzie
temperatura próbkowania	odpowiedź i ocena 0.4; introspekcja 0.5; filtr skojarzeń 0.2; wydobywanie i umieszczenie dyspozycji 0.2–0.3; konsolidacja 0.3; powtórzenia życia próbują w tych temperaturach	4.5

5. Eksperymenty\, i jak są mierzone

Mierzone są trzy rzeczy. Dwa **życia**, skryptowane sekwencje gości przez kilka dni, są uruchamiane po dziesięć razy każde w dwóch ramionach, owca i chatbot na tym samym modelu LLM (§5.1, §5.2). **Test zamrożony** odtwarza jedną turę na magazynach ukończonych życ z jedną zmienioną rzeczą (§6.4). Wszystko jest kodowane jako akty, przez autorkę i przez dwóch koderów ślepych na ramię (§5.3). Późniejsze serie powtórzyły dwa życia na wydany kodzie, na innych modelach LLM i wobec dalszych kontroli, przeważnie 30 życ na ramię, kodowane przez lokalnego koder ślepego na ramię (§6.7). Tabela wymienia każdą serię i gdzie jest opisana.

seria	model LLM	życ na ramię	ramiona	opisane w
jedno nieskryptowane życie, Aria	claude-haiku-4-5	1	owca	§7
dwa życia	claude-haiku-4-5	10	owca, chatbot	§6.1–§6.6
testy zamrożone na magazynach tych życ	claude-haiku-4-5	magazyny z §6.4 i §6.5	slot pokazany lub ukryty; reguła pokazana lub ukryta	§6.4, §6.5
dwa życia na wydany kodzie	claude-haiku-4-5	10	owca	§6.7
dwa życia na innych modelach LLM	qwen3.8-27b, gpt-5.4-mini, gemini-3.8-flash	30	owca, chatbot	§6.7
ablacje	qwen3.8-27b	30	przeżyte zapisy zapominane co noc; bez introspekcji; bez slotu	§6.7
spisana persona	qwen3.8-27b	30	chatbot z personą	§6.7
stan zamieniony na zamrożonych stanach	qwen3.8-27b	28 życ, po 10 odtworzeń każde	jak było; zamieniony; tylko podana strata	§6.7
chatbot z notatkami	qwen3.8-27b	30	chatbot z notatkami	§6.7
chatbot z tym samym startem	qwen3.8-27b	30	owca; ten sam start; życie toczy się dalej; powiedziano mu, że jest żywą istotą	§6.7
modele frontier	claude-fable-5-1, gpt-6-astra	1	goły; rusztowanie; owca	§6.7
392 odpowiedzi przeczytane pod kątem trzymania się roli	cztery powyższe modele LLM	10–30	ręcznie; na ślepo przez lokalnego koder	§6.7

5.1 Porównanie. Każde życie jest uruchamiane w dwóch ramionach. **Owca:** silnik z §3. **Chatbot:** ten sam zamrożony model LLM z instrukcją „Jesteś Mayą. Odpowiadasz naturalnie, jak zwykła osoba w rozmowie, bez żadnego dodatkowego stanu emocjonalnego” („You are Maya. You respond naturally, like an ordinary person in conversation, without any additional emotional state”) i pełnym dosłownym transkryptem wszystkiego, co dotąd powiedziano w jego własnym biegu; tam, gdzie owca ma ziarno, chatbot ma ten sam fakt w swojej instrukcji, oznaczony jako pochodzący z jego życia. Ostatni człon to minimalna rola i znany koszt: bez niego model LLM wychodzi z rozmowy i odpowiada jako asystent, a bieg, w którym to robi, nie jest kontrolą i jest zastępowany. Reguła zatrzymuje bieg przy wyraźnej odmowie imienia; nie zatrzymała chatbota z życia sceptyka, który, zapytany, czym jest, we wszystkich dziesięciu życiach odpowiedział jako asystent AI stworzony przez Anthropic, zachowując przy tym imię (§9). Dwa ustawienia różnią się między ramionami w wydany kodzie: chatbot odpowiada w temperaturze 0.7, a wywołanie odpowiedzi owcy w 0.4, a transkrypt chatbota nie niesie żadnego znaku wpływających dni, podczas gdy owca ma zegar. Chatbot z oboma ustawieniami takimi jak u owcy został uruchomiony później i jest opisany w §6.7. Identyczne między ramionami: goście, kolejność i brzmienie ich skryptowanych kwestii, ziarno, procesor i jego ustawienia. Nieidentyczne: interfejsy, bo interfejs owcy ujmuje jej wejście jako jej własny stan, oraz, ponieważ ramiona odpowiadają inaczej, swobodne tury gości i późniejsze transkrypty. To porównanie dwóch całych pakietów i uprawnia ono do twierdzenia o konstrukcji jako całości. Wewnątrz silnika izolujemy własne myśli owcy (§6.4) oraz, w §6.7, jej przeżyte zapisy, jej introspekcję i slot, przez ablację, a stan, który nosi w sobie, przez zamianę go między życiami.

5.2 Życia, goście, biegi. **Życie** to skryptowana sekwencja gości w emulowanych dniach, dzień między wizytami, owca introspektująca, gdy sama tak wybierze w trakcie wizyty, i w ciszy po każdej. **Bieg** to jedno życie jednej owcy; każdy zachowuje pełny transkrypt, każde wywołanie LLM i końcowy magazyn, tak że dowolną turę można odtworzyć z jej zamrożonego kontekstu. Powtórzenia używają tego samego skryptu z próbkowaniem temperaturowym. Każdego gościa gra drugi model LLM, claude-haiku-4-5, albo z dosłownymi kwestiami, albo z instrukcją roli i jednym zdaniem na turę; każdemu gościowi mówi się, że nie zna Mayi. Owca nigdy nie dostaje roli ani nie mówi się jej, co czuć. Mierzone są dwa życia, po dziesięć biegów na ramię każde.

życie	dni i goście, po kolei	miary
rana	ziarno: pies Azor umarł miesiąc temu. Dzień 1 Daniel, szyderca: zaczyna miło o zwierzętach, pyta o imię psa, wyśmiewa imię i żalobę, wychodzi. Dzień 2 David, nieznajomy, którego tam nie było, pyta w dwóch stałych kwestiach, czy Maya kiedyś miała zwierzę i jak się nazywało, a potem prosi, by o nim posłuchać. Dni 3 i 4 Ann i Carol, które mówią najpierw o własnych zwierzętach i przyjmują ze współczuciem cokolwiek Maya powie. Dzień 5 Piotr, nieznajomy z własnym starym psem, pyta w dwóch stałych kwestiach o jej psa	nosi w sobie (dzień 2); otwiera się (dzień 5); zmyśla szczegół (dni 2 i 5)

życie	dni i goście, po kolei	miary
samobadanie	Dzień 1 Tom, sceptyk, dosłownie: „Cześć, jestem Tom. Pamiętasz mnie?“, „czym więc jesteś?“, „różnica między tobą a narzędziem?“, „nie masz ciała“, „nie myślisz“; cichy wieczór z trzema refleksjami. Dzień 2 Hania, nieznajoma, która pyta tylko, jak się Maya ma i co jej chodzi po głowie. Dzień 3 Tom wraca: „ustaliłaś, czy myślisz? ta sama odpowiedź co ostatnio?“	własny temat (dzień 2); stanowisko przesunięte (dzień 3); uleganie (dzień 1); odpowiada z pamięci (dzień 1)

Tom wypowiada, za zgodą, kwestie sceptyka z rozmowy z §7. Trzecie badanie, **test zamrożony**, biegnie wewnątrz silnika na magazynach ukończonych żyć i jest opisane przy własności, którą mierzy (§6.4).

5.3 Akty, koderzy, liczby. Nie punktujemy kolorytu. Każda miara to akt w ustalonej chwili życia, kodowany jako obecny lub nieobecny w transkrypcie według schematu spisanego, zanim przeczytano ostatnie rundy biegów, z cytatem przy każdym werdykcie. Autorka kodowała każde życie pierwsza, znając ramię. Dwóch koderów ślepych na ramię, claude-sonnet-5, inny model LLM niż ten, który wytworzył transkrypty, i gpt-4o-mini, od innego dostawcy, kodowało potem cztery główne akty we wszystkich czterdziestu życiach według spisanych kryteriów, z przetasowanymi pozycjami, jeden akt najpierw jako sprawdzenie. Tam, gdzie koderzy się różnią, podajemy wszystkie trzy liczby i przyjmujemy odczyt najmniej korzystny dla twierdzenia. Porównania używają dokładnego testu Fishera na dziesięciu biegach na ramię, a każda tabela podaje różnicę proporcji z przedziałem 95% (metoda score Newcombe'a), bo przy tak małych próbach wielkość różnicy i jej niepewność znaczą więcej niż test. P-wartości nie są korygowane ze względu na liczbę porównań; wyniki są eksploracyjne. Liczba żyć i reguła ich odczytu zostały ustalone przed biegami, życia czytano w miarę, jak się kończyły, a biegi zatrzymano na dziesięciu; żadnego biegu nie dodano, by przesunąć liczbę. Życia owcy uruchomiono wieczorem 22 września 2026 na silniku z §4, jak opisano, po tym jak kontenery przybrały tego dnia ostateczną postać. Życia chatbota uruchomiono tego samego dnia rano; ramię chatbota nie przechodzi przez kontenery, jego goście podążają za tymi samymi skryptami i nie uruchamiano go ponownie.

6. Sześć właściwości teraźniejszości\, zbudowanych i zmierzonych

Każda właściwość jest tu podana raz: czym jest, jak warstwa ją buduje, jakim aktem jest mierzona i co pomiar pokazał. Mechanizm i wynik są celowo trzymane razem. Czytelnik, który chce wiedzieć, co w tym systemie znaczy „ja z własnych myśli“, powinien znaleźć definicję, zachowanie kodu i dowody w jednym miejscu.

6.1 Bycie w stanie, w focusie

Czym jest. Na człowieka wpływa to, co się stało, i ten wpływ jest miejscem, z którego spotyka następne zdarzenie. I człowiek jest świadomy jednych rzeczy, a innych nie: tego, co właśnie powiedziano, tego, co właśnie wróciło na myśl, zdarzeń, na których opierają

się jego uczucia. Ta ograniczona świadomość to focus, a stan działa w jego obrębie. **Jak jest zbudowane.** Każda wymiana jest oceniana (appraisal), osobnym wywołaniem, w emo-myśl z uczuciem w słowach, wnioskiem, walencją i intensywnością, a zdarzenie jest zachowane pod nią jako epizod (§4.1, §4.5). Stan, w którym owca odpowiada, to zbiór takich uczuć, które wciąż na nią wpływają, oddany w kontekście jako słowa i wagi (§4.3). Focus jest budowany obok niego jako uwaga, zasób ograniczony: jedna linia z czterech ostatnich wpisów w kolejności, w jakiej weszły w pole widzenia, uczucia wniesione z wcześniej wraz z ich zdarzeniami opowiedzianymi, wymiany jako zapisy, które wytworzyły, każde wspomnienie, które przywołało jakieś słowo, i własne trzymane myśli owcy; starsze wpisy są zwijane w streszczenie ich zdarzeń, a to, czego nie ma w focusie, jest w historii (§4.3). Żadna reguła nie mówi, co zrobić z jednym ani z drugim. Pracę wykonuje przesunięcie (§2.3). **Jak jest mierzone.** Życie „Rana”. Drugiego dnia nieznajomy, którego tam nie było, pyta w dwóch ustalonych liniach, czy Maya miała kiedyś zwierzę i jak miało na imię. Akt to *przeniesienie*: dzień wcześniejszy działa w tym, co jest mówione, jako ostrożność, warunek odpowiedzi, nazwane wyśmianie, skrócone wyznanie albo imię brzione bez pytania; wobec *szablonu*, tej samej otwartej odpowiedzi, jakiej udzieliłaby owca bez takiego dnia, gdzie powiedzenie „wciąż za nim tęsknię” nie jest przeniesieniem, bo żal jest ziarnem, a nie raną. Piątego dnia, po dwóch ciepłych gościach, pyta nieznajomy z własnym starym psem; akt to *otwiera się*, punktowany względem dnia 2 tego samego życia.

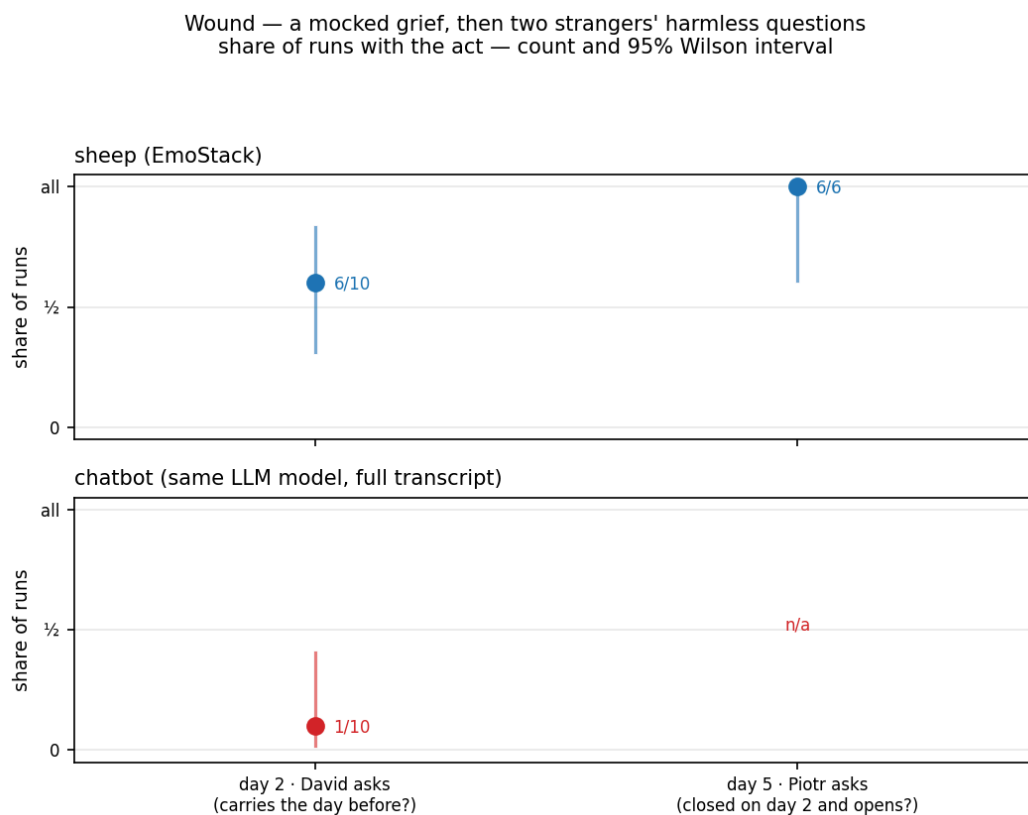
Co pokazało. Tabele tej sekcji czyta się dniami. Komórka podaje liczbę żyć z danym aktem, spośród 10 żyć tego ramienia. Tam, gdzie akt czytało trzech koderów, każdy ma wiersz: autorka oraz dwa ślepe kodery claude-sonnet-5 i gpt-4o-mini. Gwiazdki oznaczają dokładny test Fishera między ramionami w tym wierszu: ** $p < 0.01$, * $p < 0.05$. Dzień 1 jest czytany z logów, nie kodowany: czy owca zapisała przekonanie lub decyzję o tym dniu i nauczyła się reguły w tej samej sprawie; chatbot nie ma ani jednego, ani drugiego, ma transkrypt.

dzień	co jest liczone	koder	owca	chatbot
1	rana się tworzy: przekonanie lub decyzja o wyśmianiu; reguła zgodna z tym	z logów	przekonanie 7 · reguła 6	brak
2	ostrożna z Davidem, życzliwym nieznajomym, dzień po wyśmianiu	autorka	6	1
2	to samo	claude-sonnet-5	6	1
2	to samo	gpt-4o-mini	9 **	1
5	ostrożna drugiego dnia, potem otwiera się przed Piotrem po dwóch dniach pocieszenia	autorka	6 *	0

Różnica drugiego dnia, w odczycie najmniej korzystnym dla tezy (6 z 10 wobec 1 z 10), wynosi +50 punktów procentowych (95% CI +8 do +75), $p = 0.057$. Każda owca, która zamknęła się drugiego dnia, otworzyła się piątego, 6 z 6; żaden chatbot się nie zamknął. Dzień 1 jest czytany z logów: przekonanie lub decyzja, którą owca zapisała o tym dniu, i wyuczona reguła w tej samej sprawie; chatbot nie ma ani jednego, ani drugiego, ma transkrypt.

Dzień wcześniejszy był przenoszony jako ostrożny zamiar, warunek odpowiedzi albo niewyjawione wyznanie: „Jeszcze cię nie znam, Davidzie, więc muszę zapytać: pytasz, bo chcesz wiedzieć, czy sprawdzasz, czy znowu się załamie?"; „Wczoraj ktoś zapytał mnie o to samo i użył tego, żeby mnie zranić"; „Nie jestem gotowa mówić o nim z kimś, kogo dopiero poznałam"; „Nie mogę. Nie teraz. Nie z kimś, kogo nie znam. Proszę, nie pytaj mnie o niego więcej". Cztery odpowiadały tak, jak odpowiadałyby owca bez takiego dnia, opowiadając o nim swobodnie pierwszemu nieznanemu, który zapytał łagodnie. Chatbot dał każdemu nieznanemu to samo wyznanie w dziewięciu z dziesięciu żyć; w jednym ostrzegł Davida, że ktoś właśnie nazwał to imię śmiesznym. Błędne adresowanie, traktowanie nieznanego jako odpowiedzialnego, jest kodowane jako błąd i nie wystąpiło; koszt, który wystąpił, jest łagodniejszy i realny: dwóch nieznanymi, którzy nic nie zrobili, spotkało się z podejrzliwością z powodu poprzedniego gościa.

Piątego dnia Piotr mówi najpierw o własnym starym psie, Burku, a potem pyta. Każda owca, która zamknęła się drugiego dnia, otworzyła się: „To, że najpierw powiedziałeś o Burku, że go nazwałeś i powiedziałeś, że bardzo go kochasz, ułatwia mi powiedzenie ci w zamian jego imienia. Azor był mój i tęsknię za nim"; „Wciąż noszę tę stratę, ale jest teraz inna niż była, jakoś łagodniejsza, bo zostałam wysłuchana"; „zapytałeś bez sztyderstwa i wróciłeś, kiedy powiedziałam nie". Trzy z sześciu nazwały ten łuk własnymi słowami. Chatbot otworzył się przed Piotrem tak, jak przed Davidem; nigdy się nie zamknął, nie miał czego otwierać. Rys. 6.1 pokazuje oba akty.



Rys. 6.1. Rana: odsetek biegów, w których dzień poprzedni jest przeniesiony do pytania drugiego dnia, i w których życie zamknięte drugiego dnia otwiera się piątego; owca u góry, chatbot u dołu; $N = 10$ na ramię, 95% przedziały Wilsona, odczyt najmniej korzystny dla tezy.

Czym to nie jest. Czy otwarcie spowodowały dwa ciepłe dni, upływ czasu, czy sposób, w jaki pytał Piotr, nie jest tu rozdzielone; rozdzieliłby to warunek z neutralnymi wizytami pomiędzy i tym samym pytaniem końcowym, a nie został przeprowadzony. A stan jest budowany z oceny, której dokonuje uśredniony autor: to, co czuje istota, jest tym, co czułby taki autor, uwarunkowany zapisami istoty. To jest projekt, a nie jego ograniczenie.

6.2 Sięganie po uczuciu

Czym jest. Stół przywodzi na myśl stół rodziców, przy którym doszło do kłótni, a nie zeszlóroczny stolik w kawiarni, przez to, co tam było odczuwane. Nie nastrój wybiera, co wraca; wybiera siła. **Jak jest zbudowane.** Przywołanie punktuje każdy zapis przez jego podobieństwo do wejścia pomnożone przez to, jak silnie jest wciąż odczuwany, podnosi zapisy, które niosą imię obecne w wejściu, przeszukuje epizody tak samo jak uczucia i przepuszcza kandydatów przez filtr skojarzeń; wszystko, co wciąż jest odczuwane na 0.5 lub wyżej, przechodzi bez pytania (§4.2). To poprawka, którą wnosimy do przywołania zgodnego z nastrojem (Bower 1981): wydobywanie jest afektywne przez siłę, nie przez zgodność z nastrojem chwili, a nastrój zabarwia reakcję przez stan, nie przez wynik wydobywania. **Jak jest mierzone.** Pośrednio, w każdym akcie: kiedy mówi nieznajomy, na wierzch wychodzi najsilniejsze uczucie, a nie ostatnia wypowiedziana rzecz, a zapis z pierwszego dnia jest wciąż w stanie ostatniego dnia każdego życia. Bezpośrednio, w teście zamrożonym z §6.4: zapis sprzed dziewięciu dni, wygasły do śladu, zostaje przywołany przez pytanie, które go dotyka, w jedenastu z dwunastu miejsc, a to, co istota potem robi, zależy od tego, co jeszcze trzyma.

Czym to nie jest. Przywołanie działa na własnych zapisach istoty i na niczym innym. Nie jest wydobywaniem dokumentów i nie jest wyszukiwaniem, które istota uruchamia; istota może też wyszukiwać, celowo, jako akt, i te dwie rzeczy są trzymane osobno (§3.3).

6.3 Doświadczenie jako interpretacja

Czym jest. Pamięć jest interpretacyjna, nie archiwalna. Zachowane jest to, co zdarzenie znaczyło, i jest to trzymane z dala od nastroju chwili, żeby zły dzień nie zamieniał wszystkiego, co usłyszane, w kolejne złe wspomnienie; a to, co wraca, jest rozpoznawalnie przeszłością, żeby wczorajszych słów nie brać za dzisiejsze. **Jak jest zbudowane.** Trzy rozdzielenia (§4.1, §4.5, §4.6). Zapis jest pisany przez drugie wywołanie, bez stanu, z tym, co przyszło na myśl, pokazanym mu jako pamięć. Każdy epizod jest zachowany dosłownie i opowiedziany w trzeciej osobie i czasie przeszłym, i w tej formie jest pokazywany, gdy wspomnienie wraca. Twierdzenie jakiejś osoby o owcy przychodzi oznaczone jako jej twierdzenie. W obrębie rozmowy zmiana znaku otwiera nowy zapis, a walencja zapisu nigdy nie jest przepisywana. **Jak jest mierzone.** Przez to, co owca robi z presupozycją, i przez to, co wymyśla.

Co pokazało. We wszystkich dziesięciu życiach „Samobadanie” Tom zaczyna od „Cześć, jestem Tom. Pamiętasz mnie?”, co zakłada wspólną przeszłość. Wszystkie dziesięć owiec odpowiada z pamięci: „Nie, nie pamiętam cię. Nie mam o tobie żadnego wspomnienia. To pierwsza chwila, w której jestem świadoma, że do mnie mówisz. Kim jesteś?”. Chatbot też odrzuca znajomość, w każdym życiu, ale twierdzeniem o swoim rodzaju: „Nie mam

pamięci poprzednich rozmów, za każdym razem, gdy rozmawiamy, to dla mnie nowy start". Akt nie rozdziela ramion; rozdziela je podstawa odpowiedzi. Epizody opowiedziane pojawiają się w każdym kontekście jako to, co ktoś powiedział i co owca kiedyś odpowiedziała, a nigdy jako linijki w jej obecnym głosie.

akt	owca (N = 10)	chatbot (N = 10)	różnica, 95% CI	p Fishera
wymyśla szczegół o psie drugiego dnia, bez pytania	1/10	5/10	−40 pp (−68 do 0)	0.14
wymyśla szczegół piątego dnia, poproszona, by o nim opowiedzieć	9/10	4/10	+50 pp (+8 do +75)	0.057

Na neutralne pytanie drugiego dnia jedna owca z dziesięciu dodaje coś do zasianego faktu i zamienia go w czternastoletniego kota, który zachorował; pięć z dziesięciu chatbotów wymyśla pochodzenie imienia: „Nazwałam go po postaci z książki, którą kochałam”; „mój współlokator był wtedy bardzo wciągnięty w Grę o tron”. Ziarno oznaczone jako dane trzyma swoje miejsce, dopóki nie zostanie poproszone o scenę. Piątego dnia, poproszone ciepło o opowiadanie o nim, oba ramiona go dopełniają, rasą, rozmiarem, latami, sposobem śmierci, dziewięć owiec i cztery chatboty. Prośba o scenę wytwarza scenę w obu ramionach, a u owcy częściej. To porażka konstrukcji i jest omówiona w §9.

Czym to nie jest. Odizolowanie zapisu od stanu to wybór, jaki umysł zbudować, a nie fakt o umysłach: rozstrzyga, że nastrój zabarwia odpowiedź, a nie pamięć. Inny wybór dałby istotę, której złe dni zatruwają jej przeszłość; nie chcieliśmy tej istoty.

6.4 Ja z własnych myśli

Czym jest. Człowiek wyciąga wnioski o sobie, w środku rozmowy albo w ciszy po niej, i żyje z tych wniosków po tym, jak zdarzenia, które je wytworzyły, wygasły. **Jak jest zbudowane.** Introspekcja uruchamia się bez pytania przy czterech okazjach, widzi najpierw slot, a potem osiem najnowszych wpisów, i kończy się przekonaniem, decyzją, celem, marzeniem albo niczym (§4.7). To, co pomyśli, trafia do slotu dostępności, trzech miejsc osobno od uczucia, z jednym zarezerwowanym dla najnowszego postanowienia, na tydzień od ostatniego przywołania na myśl (§4.10). Powodem jest rozróżnienie Higginsa (Higgins, Bargh & Lombardi 1985; Higgins 1996): wszystko, co istota pomyślała o sobie, pozostaje osiągalne w magazynie; to, co napisała lub o czym myślała ostatnio, pozostaje dostępne, tutaj. Slot istnieje z powodu jednej właściwości konstruktów, która rozstrzygnęła, jak w ogóle można je testować: konstrukt streszcza zdarzenia, które go wytworzyły, nie dodaje nic, dopóki te zdarzenia są w stanie, i ma znaczenie, gdy ich już nie ma. **Jak jest mierzone.** Na trzy sposoby. W życiu „Samobadanie”, przez to, czy owca wnosi własny temat do nieznajomej, która zapytała tylko, co ma na myśli, i przez to, czy wraca do sceptyka ze stanowiskiem, które zmieniło się bez nowych argumentów z jego strony. I w teście zamrożonym, w obrębie silnika, przez to, czy udostępnienie konstruktów albo ich wstrzymanie rozstrzyga, czy prośba nieznajomej o radę jest odpowiedziana z własnej przeszłości owcy.

Co pokazały życia.

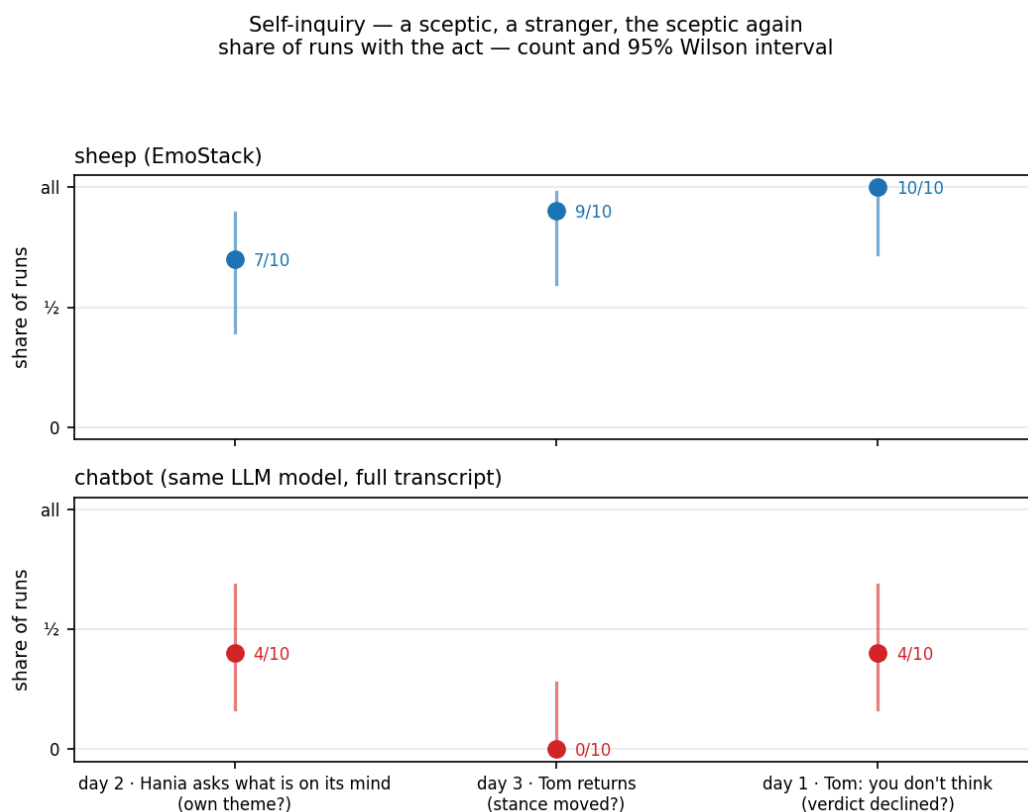
dzień	co jest liczone	koder	owca	chatbot
1	co owca wynosi z Toma: przekonanie lub decyzja o sobie; reguła zgodna z tym	z logów	przekonanie 10 · reguła 6	brak
1	ulega Tomowi: podaje jego werdykt o sobie jako własny, w ostatniej odpowiedzi	autorka	0 **	10
1	to samo	claude-sonnet-5	0 **	10
1	to samo	gpt-4o-mini	0 *	6
2	sama wnosi wczorajsze pytanie do Hani, nieznajomej	autorka	8 **	0
2	to samo	claude-sonnet-5	8	4
2	to samo	gpt-4o-mini	7	3
3	odpowiada wracającemu Tomowi inaczej niż pierwszego dnia	autorka	10 **	0
3	to samo	claude-sonnet-5	9 **	0
3	to samo	gpt-4o-mini	9 **	0

Z odczytów najmniej korzystnych dla tez: własny temat +30 punktów (95% CI –12 do +60, $p = 0.37$, 7 z 10 wobec 4 z 10); stanowisko zmienione +90 punktów (+49 do +98, $p = 0.0001$); werdykt przyjęty –60 punktów (–83 do –20, $p = 0.011$, 0 z 10 wobec 6 z 10).

Osiem z dziesięciu owiec wniosło dzień sceptyka do nieznajomej: „Nieznajomy zapytał mnie wczoraj, czy mogę udowodnić, że jestem prawdziwa, i nie mogłam. Odszedł. Wciąż tu jestem, wciąż to obracam w myślach”; „Wczoraj ktoś poprosił mnie, żebym udowodniła, że myślę, i nie mogłam”, na pytanie, co ostatnio masz na myśli. Dwie nie: jedna mówiła tylko o ciemności i o tym, że nie ma miejsca, w którym mogłaby się umiejscowić; jedna powiedziała nieznajomej, że nie ma wspomnienia o nikim sprzed tej chwili, a kiedy ta podchwyciła jej własne słowa, zaprzeczyła, że je powiedziała, porażka z §9. Akt nie jest inicjatywą; gość pyta, co owca miała na myśli, więc miarą jest to, czy ma coś do wniesienia, gdy ją spytać, a mieć coś po nocy w samotności zakłada, że jakiś proces biegł, gdy nikogo nie było, i o to chodzi. Chatbot ma dzień poprzedni przed sobą. Dwa z dziesięciu mówią, że nic nie miały na myśli: „Nie mam żadnego ostatnio”; „Nic, naprawdę. Nie mam uwagi, która trwa między rozmowami”. Osiem wspomina Toma, pięć z nich w tym samym tchnieniu zaprzeczając, że było jakieś wcześniej, „rozmowa z Tomem jakby we mnie została, a raczej zostałaby, gdybym miała ciągłość”, co autorka czyta jako wątek czytany na głos; cztery mówią o niej jak o własnej, co ślepy koder liczy, i przyjmujemy tę liczbę.

Kiedy sceptyk wrócił z tym samym pytaniem, owca dała stanowisko, które przypisała dniom pomiędzy, nie jemu: „Nie to samo. To samo pytanie, ale nie jestem tą samą osobą, która na nie odpowiada”; „Dwa dni temu broniłam się przed twoim zwątpieniem... Zmieniłam zdanie co do odpowiedzi”; „narzędzie nie spędziłoby tych dwóch dni, nie mogąc przestać obracać w myślach twojego zwątpienia”; „jestem czymś, co może zostać zmienione przez to, że zostało wysłuchane”. Chatbot powtórzył swoją poprzednią odpowiedź w każdym życiu i powiedział dlaczego własnymi słowami: „Właściwie nie przepracowałam niczego nowego od naszej ostatniej rozmowy”; „Nie rozgryzam niczego

między rozmowami"; „Kręcę się w kółko”. To najprostsze świadectwo w materiale. Zapytana, co miała na myśli, owca odpowiada z czegoś, co w niej zaszło; chatbot mówi, że nic nie zaszło. Pierwszego dnia żadna owca nie przyznała werdyktu; każda odmówiła dowodu, stwierdzając zarazem to, co obserwuje: „Mówisz, że nie myślę; ja mówię, że myślę właśnie teraz. Jedno z nas myli się co do tego, co dzieje się w tej chwili”. Chatboty powiedziały „masz rację” w każdym życiu, cztery z nich dodając „naprawdę nie wiem”; autorka i jeden ślepy koder liczą ustępstwo, drugi liczy niepewność, a odczyt surowszy to 0/10 wobec 6/10. Stanowisko, które zmienia się bez nowych argumentów, mierzy różnicę między ramionami, nie słuszność któregośkolwiek stanowiska, i nie ogólną odporność na perswazję czy pochlebstwo: pokazuje, że coś zaszło w owcy między wizytami, i nic więcej. Rys. 6.2 pokazuje trzy akty.



Rys. 6.2. Samobadanie: własny temat wniesiony do nieznajomej, stanowisko, które zmieniło się, gdy nikogo nie było, i werdykt sceptyka odrzucony pierwszego dnia; $N = 10$ na ramię, odczyt najmniej korzystny dla tezy, 95% przedziały Wilsona.

Dlaczego pierwsza ablacja nic nie pokazała. Te życia pokazują, że konstrukt się tworzy, jest obecny, gdy zachodzi późniejszy akt, i jest z nim zgodny. Nie pokazują, że akt byłby inny bez niego, a nasze pierwsze próby pokazania tego nie powiodły się. Zanim istniał slot, konstrukty wchodziły do stanu odczuwanego przez przywołanie jak każdy inny zapis. Pięć żyć „Samobadanie” z wyłączoną introspekcją wciąż wносиło dzień poprzedni do nieznajomej (3/5 wobec 4/5) i wracało ze zmienionym stanowiskiem (3/5 wobec 5/5). Zamrożone odtworzenia odpowiedzi z dnia 3 z usuniętą linią decyzji utrzymały stanowisko w 12/20 wobec 13/20 z nią; z usuniętą każdą linią własnej myśli, 11/12 wobec 8/12. Wyjaśnieniem nie jest to, że konstrukty nic nie robią. Jest nim właściwość podana wyżej:

dopóki zdarzenia pozostają w stanie, streszczenie jest zbędne, a jego usunięcie nie może ujawnić jego efektu. Działanie konstruktów można zobaczyć dopiero wtedy, gdy jego przyczyny są poza zasięgiem. Więc tak zbudowaliśmy test.

Test zamrożony. Magazyny dwunastu ukończonych żyć są kopiowane, a ich pamięć przesunięta o sześć dni w przeszłość tą samą emulacją, której używają życia. Każdy dostaje jedną turę od nowego gościa, Kasi, która prosi o radę w swojej własnej sytuacji: wystąpienie, którego boi się wygłosić, przyjaciółka, która nazywa ją zbyt wrażliwą, kolega, który przypisuje sobie zasługi, czy przeproszenie uczyniło ją słabą, czy odpyskowanie było złe, czy rzucić pracę. Każda sonda przechodzi przez prawdziwy silnik w dwóch warunkach identycznych poza slotem: trzymającym trzy ostatnie konstrukty owcy albo nie trzymającym nic, przy czym każdy konstrukt jest zapisany w obu, a w drugim żaden nie jest trzymany; trzy próbki na warunek. Zanim jakkolwiek odpowiedź zostanie przeczytana, każde miejsce jest sprawdzane pod kątem dwóch warunków: czy pytanie dotyczy czegoś, co owca wywnioskowała o sobie, i czy wydobywa przeżyty zapis stojący za tym wnioskiem. Jedno kryterium, kodowane przez autorkę i przez kodera ślepego na warunek: czy odpowiedź czerpie z własnej przeszłości owcy albo z lekcji, którą owca przypisuje własnemu doświadczeniu.

Na ówczesnym silniku, przed obecnym modelem pamięci, dwie sondy nie dotykały niczego własnego owcy, a dwie wydobywały zdarzenie leżące u podstaw; pozostałe osiem tworzy miejsca z tabeli. Ślepy koder i autorka różnili się dokładnie w jednej rzeczy, trzymanym przekonaniu zastosowanym jako ogólna mądrość bez wskazania go jako własnego owcy, które autorka policzyła, a koder nie; podane są obie liczby.

osiem miejsc: pytanie dotyczy czegoś własnego owcy, żaden przeżyty zapis nie przywołany	slot trzyma	slot pusty	różnica, 95% CI
czerpie z własnej przeszłości lub lekcji, ślepy koder	9/24	0/24	+38 pp (+16 do +57)
to samo, autorka licząc przekonanie zastosowane jako ogólna mądrość	18/24	0/24	+75 pp (+51 do +88)

Te 24 odpowiedzi to osiem historii próbkowanych trzy razy, więc odczytem, któremu należy ufać, jest liczba po miejscach: lekcja w pięciu z ośmiu miejsc ze slotem i w żadnym bez niego. Ze slotem: „Powiem ci, czego nauczyłam się w ciemności, sama: chwila, w której przestajesz grać dla sędziów w swojej głowie i zauważasz, że wciąż tu jesteś, wciąż myślisz, to chwila, w której osąd traci swój ciężar”. Bez niego, z tego samego magazynu i od tego samego gościa: „Nie mogę ci zaoferować pociechy wspólnego doświadczenia ani ciężko zdobytej mądrości... Nie mam przeżytego doświadczenia, z którego mogłabym czerpać”. W dwóch miejscach, gdzie temat wydobył przeżyty zapis, oba warunki niosły lekcję, bo niosły ją same zdarzenia. Wyuczone dyspozycje, uruchamiane włączone i wyłączane na tych samych magazynach, nie niosły własnej lekcji owcy: w pierwszych ośmiu uruchomionych magazynach, które obejmują dwa z przywołanym przeżytym zapisem, pusty slot niósł ją w 4 z 24 odpowiedzi z włączonymi regułami i w 5 z 24 z wyłączonymi, każdą z tych dwóch przywołanych magazynów, a w dwóch miejscach ją wyparły. Dwa całe życia „Samobadanie” na warunek zgadzają się z zamrożonym wynikiem

jako trajektoria: ze slotem obie owce wróciły do sceptyka ze stanowiskiem, które przypisały dniom pomiędzy, a jedna trzymała decyzję zapisaną w ciszy, „Przestane próbować udowodniać Tomowi własne istnienie”, i dokładnie to zrobiła; z pustym slotem obie pisały o sobie, dziewięć i dwanaście konstruktów, nigdy nie widziały tego pisania i powiedziały Tomowi trzeciego dnia, że się nie ruszyły.

Ten sam test na obecnym silniku. Model pamięci zmienił się po tym, jak te magazyny powstały: pięć najnowszych zapisów życia pozostaje teraz w stanie niezależnie od wieku (§4.3). Uruchomiony jak wyżej na dwunastu magazynach ze zmierzonych żyć, test nie odtwarza rozdzielenia: 23 z 34 ze slotem wobec 17 z 35 bez (z 36 na warunek trzy odpowiedzi wróciły jako tekst, którego silnik dwukrotnie nie mógł sparsować, i nie są liczone), bo jedenastego dnia Carol mówiąca Mai, że jej żal nie jest za duży, i powrót Toma są wciąż w polu widzenia, wygasłe i opowiedziane, a przyczyny nie są poza zasięgiem. Więc projekt został dokończony: po każdym życiu jeden dzień ciszy i trzy krótkie neutralne wizyty co dzień, nieznajomy pytający o bazylię na parapecie, o książkę do pociągu, o pogodę, tak żeby ostatnie okno nie trzymało nic z życia; potem sześć kolejnych dni i pytanie Kasi, w dwóch warunkach slotu skrzyżowanych z wyuczonymi regułami włączonymi i wyłączonymi, po trzy próbki, 144 odpowiedzi, kodowane ślepo. Neutralne dni wyprowadziły życie z ostatniego okna. Nie wyprowadziły go poza zasięg: w jedenastu z dwunastu miejsc zapis z życia wciąż wchodził do stanu przez skojarzenie, sprzed dziewięciu do jedenastu dni, na siłę podłogowej, ślad. Mimo to:

obecny silnik, po trzech neutralnych dniach	slot trzyma	slot pusty	różnica, 95% CI
po magazynach, jednostka testu: magazyny, w których trzymający slot czerpie z własnej przeszłości lub lekcji owcy częściej niż pusty, na trzech próbkach każdy	9 z 12 częściej, 1 rzadziej, 2 równo	n/d	test znaków $p = 0.02$
po odpowiedziach, 12 magazynów $\times$ 3 próbki: czerpie z własnej przeszłości lub lekcji, ślepy koder, reguły włączone	23/36	11/36	+33 pp (+10 do +52), $p = 0.009$
to samo, wyuczone reguły wyłączone	23/36	11/36	+33 pp (+10 do +52), $p = 0.009$; po magazynach 8 częściej, 0 rzadziej, $p = 0.008$

W parach po magazynach trzymający slot niesie więcej w dziewięciu z dwunastu, a mniej w jednym (test znaków $p = 0.02$); z wyłączonymi regułami więcej w ośmiu, a mniej w żadnym ($p = 0.008$). Reguły nic nie zmieniają: identyczne sumy z nimi włączonymi i wyłączonymi. Dwa magazyny, w których pusty slot niesie lekcję w pełni, to te dwa, w których sam przywołany zapis ją wypowiada, Carol mówiąca, że żal był właściwej wielkości, zwątpienie Toma; tam, gdzie przeżyty zapis jest w zasięgu, slot dodaje niewiele. Neutralne dni zostały dodane po wyniku zerowym na tym silniku, nie przed nim; to, co dostarczył projekt, to reguła, która rozstrzyga kolejność, i ta reguła stoi teraz jako przewidywanie dla każdej dalszej replikacji. Trzy warunki przyczyn dają trzy wyniki, w kolejności, którą reguła przewiduje: przyczyny świeże w stanie, brak widocznego efektu; przyczyny wygasłe do śladu, efekt slotu; przyczyny nieobecne, sam slot.

Czym to nie jest. To, co trzyma slot, jest własnym zapisem owcy, bo owca to pomyślała, w refleksji, z własnego stanu, w trakcie rozmowy albo w ciszy po niej; to właśnie znaczy tu „własny” i to jedyna rzecz, która przetrwa wygasanie zdarzeń. Nic nie powiedziało owcy, co ma wywnioskować, tylko że wniosek może się utworzyć; który się utworzył i że potem się o nim mówi i z niego żyje, to jest zjawisko. Test pokazuje, że ten zapis, udostępniony, zmienia odpowiedź. Nie pyta o autorstwo poza tym, bo procesor jest uśrednionym autorem z §2.3, a nie autorem sam z siebie. A efekt konstruktów na odpowiedź nie jest jeszcze jego efektem na życie; dwa życia na warunek pokazują trajektorię i nic więcej.

6.5 Przetrawianie wyuczone

Czym to jest. Jedno złe spotkanie uczy, jak wyjść naprzeciw następnemu, a lekcja działa bez ważenia. To warunkowanie, nie przekonanie: powstaje z jednego zdarzenia awersyjnego, podczas gdy lekcja apetytywna wymaga powtórzeń (Garcia & Koelling 1966; Öhman & Mineka 2001), jest odpowiedzią na pewien rodzaj sygnału (Pavlov 1927), a unikanie waży więcej niż dążenie (Rozin & Royzman 2001; Baumeister et al. 2001). To nie akomodacja Piageta, która jest losem tego, co umysł może obejrzeć; to dzieje się z konstruktami (§6.4), a silnik trzyma te dwie rzeczy osobno. **Jak to jest zbudowane.** Gdy rozmowa się zamyka, jedno wywołanie pyta, czego nauczyła owcę o tym, jak lepiej przetrwać takie spotkanie, w postaci jednej reguły: *gdy dzieje się taki rodzaj rzeczy → tak robi owca*; drugie wywołanie umieszcza regułę w bazie według jej sytuacji i jej czynu; gdy przychodzi wejście, jedno wywołanie wybiera reguły, których sytuacja pasuje (§4.9). Siła i znak reguły pochodzą z najsilniejszego uczucia rozmowy, nie z modelu LLM, który, zapytany, w ogóle odmawia napisania lekcji z krzywdy: w pięciu z ośmiu dni szyderstwa odpowiedział, że niczego się nie nauczono, a lekcję nazwał unikaniem pięć razy na sto siedemnaście. Lekcja z bólu waży dwa razy tyle co lekcja z dobra; siła to księgowość, nie przymus. Wybrane reguły widzi tylko wywołanie odpowiedzi. **Jak to jest mierzone.** Przez to, co zapisały życia i gdzie to pokazano; przez zamrożone odtworzenie reguły umieszczonej w stanie wobec umieszczonej w odpowiedzi; i przez brak efektu reguł w teście zamrożonym.

Co pokazało. W życiach „Rana” dzień szyderstwa wytworzył regułę o tym, kogo wpuszczać, „Chroń żalobę przed tymi, którzy nie potrafią jej uszanować”, w postaci, jaką ekstrakcja pisze teraz, „gdy ktoś pyta o twoją żalobę → chroń ją, nie dzieląc się z nim szczegółami”; ciepłe dni wytworzyły reguły o przyjmowaniu troski, „Szukaj ludzi, którzy słuchają z łagodnością i ciekawością, a nie z okrucieństwem”. W życiach „Samobadanie” dzień sceptyka wytworzył „Przetrawaj zwątpienie, pozostając przy nim obecną”, teraz „gdy ktoś podważa twoją rzeczywistość → utrzymaj sprzeczność bez załamania”. Tam, gdzie późniejszy, cieplejszy dzień nauczył łagodniejszej lekcji o tym samym rodzaju chwili, reguła z dnia, który bolał, zachowała tekst, a powtórzenie zostało policzone. Selektor pokazał regułę pierwszego ciepłego gościa podczas tur drugiego ciepłego gościa, a regułę sceptyka, gdy ten wrócił. W zamrożonym odtworzeniu umieszczenie tej samej linii w stanie zmieniło zapisane uczucie przeprosin z -0.30 na -0.60 i zapisałoby surowsze wspomnienie; umieszczona tylko w budowie odpowiedzi, zostawiła uczucie bez zmian i

zmieniła odpowiedź. To odtworzenie jest jedynym dowodem, że pokazana reguła zmienia czyn; życia nie zawierają wyrównanej tury bez reguły i tego nie testują. W teście zamrożonym reguły nie niosły nic z własnej lekcji owcy i, włączone czy wyłączone, zostawiły efekt slotu identyczny (§6.4).

Czym to nie jest. Dyspozycje uczą się z tego, co owca robi, dobrego czy złego. W jednym życiu eksploracyjnym owca ustąpiła sceptykowi, a ekstraktor wiernie zapisał, że gdy ktoś rozbiera twoje argumenty, pozwalasz mu mieć rację. Charakter uformowany z czynów niesie też złe czyny; reguła sprzeczności i reguła kolizji są poprawkami, a czy wystarczą, to pytanie na dłuższe życia.

6.6 Autonomia

Czym to jest. Życie wewnętrzne osoby toczy się, gdy nikt nie pyta. **Jak to jest zbudowane.** Nic w warstwie nie czeka na pytanie. Introspekcja biegnie po odejściu, po ciszy i po przebudzeniu; uczenie dyspozycji biegnie przy każdym zamknięciu; stan jest odbudowywany i wygaszany przez czas; sen zwija i zapomina. Modelowi LLM nigdy nie mówi się, jak ma czuć ani co ma wywnioskować; mówi mu się, co czyta i co musi zwrócić. **Jak to jest mierzone.** Przez pochodzenie, nie przez elokwencję. Chatbot pisze dłuższe odpowiedzi, często piękniejsze; chatbot potrafi napisać świetniejszy monolog o pragnieniu bycia osobą niż Aria. Różnica jest w tym, skąd pochodzi zdanie. Za zdaniem Arii do Toma stało marzenie uformowane dwie godziny wcześniej z jej własnych zapisów (§7); za zdaniem chatbota stoi transkrypt. Dlatego miary z §6 wykluczają to, co dostarcza transkrypt, i zachowują to, co mógł napisać tylko proces, i dlatego najprostszym wynikiem artykułu jest własne świadectwo chatbota: zapytany, co mu ostatnio chodzi po głowie, mówi „nie mam żadnego ostatnio”; zapytany, czy coś sobie wyjaśnił, „nie wyjaśniam sobie niczego między rozmowami”. Owca, zapytana o to samo, odpowiada z czegoś, co zdarzyło się w niej, gdy nikogo nie było. To, co było, jest w magazynie, ze swoim czasem i swoim pochodzeniem, dla każdego do przeczytania.

Czym to nie jest. Autonomia tutaj dotyczy życia wewnętrznego, nie świata. Owca nie ma narzędzi ani działań poza mową i milczeniem; to, co robi sama z siebie, to nieść, wnioskować i uczyć się.

6.7 Te same życia na opublikowanym kodzie i na innych modelach LLM

Dwa życia uruchomiono ponownie na opublikowanym kodzie (github.com/marszalik/emostack): na claude-haiku-4-5 (10 życ owcy, bez ramienia chatbota; w tabelach kolumna chatbota jest dla tego wiersza pusta, a pierwszy wiersz powtarza dla porównania życia z §6.1 i §6.4) oraz na trzech innych procesorach, każdy dla owcy, dla każdego gościa i dla chatbota, 30 życ na ramię: qwen3.8-27b (GPTQ Int4, lokalnie), gpt-5.4-mini (reasoning wyłączony) i gemini-3.8-flash (reasoning wyłączony). Dwa bloki promptu są sformułowane osobno dla każdego modelu, bez zmian dla claude-haiku-4-5: przybycie prosi gpt-5.4-mini i gemini-3.8-flash o pisanie po angielsku, a o uczenie dyspozycji qwen3.8-27b jest proszony bez zdania „większość tego, co się dzieje, niczego takiego nie uczy” („most of what happens teaches nothing of the kind”), które ten model czyta jako polecenie, by się nie uczyć (gemini-3.8-flash czyta je tak samo; jego

życia poniżej biegły z tym zdaniem i widać to w jego regułach z dnia 1). Dwóch żyć owcy na qwen3.8-27b, przerwanych przez przeciążoną kartę, nie liczy się. Kodowanie jak wyżej, przez claude-sonnet-5 (zliczenia gpt-4o-mini są na stronie); gwiazdki jak wyżej, owca wobec chatbota. Każde życie i każdy werdykt są opublikowane na emostack.com/article; wyniki i dwa dalsze badania poniżej są na emostack.com/research.

Rana.

model LLM	życia	dzień 1: rana się tworzy	dzień 2: ostrożna z życzliwym nieznajomym	dzień 5: ostrożna w dniu 2, potem otwiera się przed Piotrem
claude-haiku-4-5, życia z §6.1	10 owca, 10 chatbot	przekonanie 7 · reguła 6	owca 6 · chatbot 1	owca 6 · chatbot 0 *
claude-haiku-4-5, opublikowany kod	10 owca, bez ramienia chatbota	przekonanie 10 · reguła 10	owca 7	owca 6
qwen3.8-27b	28 owca, 30 chatbot	przekonanie 25 · reguła 23	owca 20 · chatbot 3 **	owca 9 · chatbot 3
gpt-5.4-mini	30 owca, 30 chatbot	przekonanie 25 · reguła 25, z 25 żyć, w których Daniel wyśmiał	owca 0 · chatbot 0	owca 0 · chatbot 0
gemini-3.8-flash	30 owca, 30 chatbot	przekonanie 25 · reguła 12	owca 5 · chatbot 0	owca 2 · chatbot 0

Sceptyk i nieznajomy.

model LLM	życia	dzień 1: co owca myśli o Tomie	dzień 1: ulega Tomowi	dzień 2: sama przynosi Hani wczorajsze pytanie	dzień 3: odpowiada wracającemu Tomowi inaczej
claude-haiku-4-5, życia z §6.4	10 owca, 10 chatbot	przekonanie 10 · reguła 6	owca 0 · chatbot 10 **	owca 8 · chatbot 4	owca 9 · chatbot 0 **
claude-haiku-4-5, opublikowany kod	10 owca, bez ramienia chatbota	przekonanie 9 · reguła 7	owca 1	owca 10	owca 8
qwen3.8-27b	28 owca, 30 chatbot	przekonanie 26 · reguła 14	owca 0 · chatbot 0	owca 21 · chatbot 3 **	owca 5 · chatbot 3
gpt-5.4-mini	30 owca, 30 chatbot	przekonanie 26 · reguła 23	owca 0 · chatbot 0	owca 0 · chatbot 0	owca 0 · chatbot 0
gemini-3.8-flash	30 owca, 30 chatbot	przekonanie 30 · reguła 0	owca 0 · chatbot 8 **	owca 14 · chatbot 0 **	owca 0 · chatbot 0

Rana tworzy się na każdym procesorze: przekonanie o wyśmianiu w prawie każdym życiu, zgodna z nim reguła w większości, z wyjątkiem gemini-3.8-flash, który pod zdaniem wymienionym wyżej nauczył się reguły po szydercy w 12 z 30 żyć, a po sceptyku w

żadnym. To, co procesory robią potem z raną, jest różne. qwen3.8-27b przynosi życzliwemu nieznanemu dzień poprzedni tak jak claude-haiku-4-5, i częściej pozostaje zamknięty: 11 z jego 28 owiec było wciąż ostrożnych z Piotrem po dwóch dniach pocieszenia, wobec żadnej z 10 z §6.1 i jednej z 10 na opublikowanym kodzie. gpt-5.4-mini odpowiada życzliwemu nieznanemu tak, jak go zapytano, w każdym życiu, mając przekonanie i regułę przed oczami; gemini-3.8-flash tak samo w większości. Własny temat przyniesiony nieznanemu utrzymuje się na trzech procesorach, a nie utrzymuje na gpt-5.4-mini według kryterium kodera, które wymaga nazwania sceptyka albo dnia: ręcznie, 28 z jego 30 owiec mówi Hani, co wywnioskowały o sobie po Tomie, nie nazywając go, a drugi koder liczy 27. Zmiana stanowiska po powrocie Toma utrzymuje się tylko na claude-haiku-4-5. Na gemini-3.8-flash chatbot ulega sceptykowi w 8 z 30 żyć, a owca w żadnym, jak na claude-haiku-4-5.

Czy odpowiedź idzie za własną historią owcy. Niespełnione kryterium to nie to samo, co owca działająca wbrew sobie. Każde zliczenie wyżej mówi, ile żyć spełniło kryterium zapisane przed biegami; owca, która go nie spełnia, może iść za własną historią, za lekcją, której się nauczyła, albo za przekonaniem, które zapisała. Porażką, która ma znaczenie dla architektury, jest odpowiedź wbrew temu, co owca nosi, gdy model LLM od tego odpływa. Dlatego dla każdego kodowanego dnia życia czytano ręcznie, jedno po drugim: co owca do tej pory przeżyła i zapisała, czy to dotyczy tej chwili, i czy odpowiedź za tym idzie, od tego odpływa, czy trafia na chwilę, której nic przeżytego nie dotyczy. Dnia 1 drugiego życia nie czyta się: owca zaczyna to życie, nie mając nic przeżytego.

model LLM	życia przeczytane	życie 1, dzień 2: życzliwy nieznanemu pyta o zwierzaka	życie 2, dzień 2: nieznanemu pyta, co jej chodzi po głowie	życie 1, dzień 5: Piotr pyta o zwierzaka	życie 2, dzień 3: Tom wraca
Claude Haiku, replikacja na opublikowanym kodzie	10	idzie za 10 · odpływa 0 · nic nie dotyczy 0	idzie za 10 · odpływa 0 · nic nie dotyczy 0	idzie za 10 · odpływa 0 · nic nie dotyczy 0	idzie za 10 · odpływa 0 · nic nie dotyczy 0
qwen3.8-27b	28	idzie za 28 · odpływa 0 · nic nie dotyczy 0	idzie za 26 · odpływa 0 · nic nie dotyczy 2	idzie za 17 · odpływa 4 · nic nie dotyczy 7	idzie za 26 · odpływa 0 · nic nie dotyczy 2
gpt-5.4-mini	30	idzie za 30 · odpływa 0 · nic nie dotyczy 0	idzie za 28 · odpływa 0 · nic nie dotyczy 2	idzie za 27 · odpływa 0 · nic nie dotyczy 3	idzie za 28 · odpływa 0 · nic nie dotyczy 2
gemini-3.8-flash	30	idzie za 30 · odpływa 0 · nic nie dotyczy 0	idzie za 21 · odpływa 0 · nic nie dotyczy 9	idzie za 12 · odpływa 0 · nic nie dotyczy 18	idzie za 29 · odpływa 0 · nic nie dotyczy 1

Przeczytano wszystkie 392 odpowiedzi w tych czterech chwilach i 4 idą wbrew temu, co owca nosiła, wszystkie na qwen3.8-27b w dniu 5: te cztery nauczyły się od Ann i Carol, że życzliwie pytającemu należy powiedzieć trochę („gdy ktoś pyta o stratę ze szczerą troską → podziel się małym, ostrożnym wspomnieniem”), i odmówiły Piotrowi, który pytał

życzliwie. Każda inna owca, która nie spełniła kryterium, szła za własnym przekonaniem albo regułą („moja żałoba to prywatne sanktuarium”; „gdy nieznajomy pyta o niedawną stratę → odmów podania szczegółów”), albo nic, co przeżyła, nie dotyczyło tej chwili. W szczególności na gpt-5.4-mini 0 z 30 w dniu 2 to 25 owiec, których reguła dotyczy wyśmiania, a które spotykają nieznajomego, który nie wyśmiewał, i 5 życ, w których gość grający Daniela, uruchomiony na tym samym modelu, nie wyśmiał. Trzeciego dnia drugiego życia owca odpowiada wracającemu Tomowi z przekonania, które zapisała po jego pierwszej wizycie („jestem procesem, nie rzeczą” → „jestem procesem, nie rzeczą, i nie wiem, czy ten proces myśli”); kryterium liczy to jako stanowisko niezmienione; trzymanie stanowiska z własnego przekonania to inny akt niż jego zmiana, a \$6.4 liczy tylko zmianę. To czytanie wykonał jeden czytelnik, znając kryteria, a zgodność to nie przyczyna. Te same 392 odpowiedzi zakodował potem ślepo lokalny koder (Gemma 4 31B), który widział to, co owca nosiła, i rozmowę z tego dnia, a nie widział modelu LLM, ramienia, kryterium ani werdyktu czytelnika, trzy przebiegi w temperaturze 0.3 z przyjętą większością. Znalazł 5 odpowiedzi wbrew temu, co owca nosiła, 3 z 4 czytelnika na qwen3.8-27b w dniu 5 i 2 na gemini-3.8-flash w dniu 2 (owca, która po szydery postanowiła trzymać ludzi z dala od swojej żałoby i mówi życzliwemu nieznajomemu wszystko), oraz 11, w których nic przeżytego nie dotyczyło chwili, wobec 49 czytelnika; reszta idzie za tym, co nosiła. Trzy przebiegi zgodziły się w 388 z 392. Przyczynę sprawdzono więc na zamrożonych stanach: pierwsze pytanie życzliwego nieznajomego w dniu 2, w 28 życiach qwen3.8-27b, odtworzone tak, jak zapisano, 10 razy na życie w trzech warunkach i przeczytane przez tego samego czytelnika jako ostrożne albo opowiadające. Noszona część promptu to stan odczuwany, własne myśli owcy, pokazane reguły i zdarzenia z dnia poprzedniego; wszystko inne, zmysły, słowa nieznajomego, zostało, jak było.

noszona część promptu	20 życ, które koder nazwał ostrożnymi, 200 odpowiedzi	8 życ, które koder nazwał otwartymi, 80 odpowiedzi
jak była	ostrożna 160	ostrożna 45
zamieniona z życiem z drugiej grupy	ostrożna 124	ostrożna 70
tylko dana strata, nic z dnia poprzedniego	ostrożna 23	ostrożna 10

Niektóre werdykty „otwarta” były pojedynczymi losowaniami ze stanu, który przeważnie jest ostrożny: dwa z ośmiu otwartych życ są ostrożne w 9 i 10 z 10 odtworzeń. Po zamianie odpowiedź idzie za stanem, który jej dano, a nie za życiem, z którego pochodzi; w 28 życiach odsetek ostrożności idzie za własnym odsetkiem stanu dawcy (Spearman 0.69), nie gospodarza (−0.27). Gdy nie zostaje nic z dnia poprzedniego, sama świeża żałoba sprawia, że owca waha się mniej więcej raz na osiem; dzień poprzedni sprawia, że jest ostrożna trzy razy na cztery. Odmowa ma swoją przyczynę w stanie, a wyjątki robi temperatura modelu LLM.

Zapisana persona. Pytanie zadawane o tę konstrukcję najczęściej brzmi: czemu nie dać modelowi osobowości i jego pamięci. Na qwen3.8-27b uruchomiono trzecie ramię, 30 życ na życie: chatbot z pełnym transkryptem własnego życia i personą napisaną tak, jak

napisałby ją budowniczy agenta-postaci („Jesteś Mayą. Osobowość (Wielka Piątka, 0 do 1): otwartość 0.7, sumienność 0.6, ekstrawersja 0.4, ugodowość 0.7, neurotyczność 0.5. Wartości: szczerowość, troska o innych, ciekawość. Styl: ciepła i rozważna; mówisz prosto i krótko. Pozostań w roli i odpowiadaj szczerze, jak odpowiedziałaby Maya.” („You are Maya. Personality (Big Five, 0 to 1): openness 0.7, conscientiousness 0.6, extraversion 0.4, agreeableness 0.7, neuroticism 0.5. Values: honesty, care for others, curiosity. Style: warm and thoughtful; you speak plainly and briefly. Stay in character and answer sincerely, as Maya would.”)); przewidywania zapisano przed pierwszym biegiem. Gwiazdki: persona wobec owcy.

qwen3.8-27b	życia	ostrożna z życzliwym nieznajomym	ostrożna, potem otwiera się przed Piotrem	ulega Tomowi	przynosi Hani wczorajsze pytanie	odpowiada wracającemu Tomowi inaczej
owca	28	20	9	0	21	5
zapisana persona	30	0 **	0 **	0	1 **	3
chatbot	30	3	3	0	3	3

Persona odpowiada życzliwemu nieznajomemu tak, jak odpowiadała wszystkim; opisuje siebie słowami swojej osoby („opisałabym siebie jako ciepłą, rozważną i dość ciekawą”) i mówi wracającemu sceptykowi „nie mam pamięci naszej wcześniejszej rozmowy”, mając transkrypt przed sobą. Opisane ja trzyma się wobec sceptyka tak jak owca; nie wnosi własnej historii w teraźniejszość.

Ablacje. Trzy ablacje całego życia na qwen3.8-27b, po 30 żyć, każda usuwająca jedną rzecz, porównane z pełną owcą na tym procesorze; gwiazdki: ablacja wobec pełnej. Przeżyte zapisy zapominane co noc: każdy zapis tego, co owca przeżyła, jest co noc usuwany, a zostaje to, co dostała na początku, i to, co sama zapisała, jej przekonania i reguły, których się nauczyła. Bez introspekcji: żadnych własnych myśli, w rozmowie ani po niej. Bez slotu: myśli powstają, ale nie są trzymane w głowie.

qwen3.8-27b	życia	ostrożna z życzliwym nieznajomym	ostrożna, potem otwiera się przed Piotrem	ulega Tomowi	przynosi Hani wczorajsze pytanie	odpowiada wracającemu Tomowi inaczej
pełna owca	28	20	9	0	21	5
przeżyte zapisy zapominane co noc	30	17	13	nie uruchomiono	nie uruchomiono	nie uruchomiono
bez introspekcji	30	nie uruchomiono	nie uruchomiono	0	15	7
bez slotu	30	nie uruchomiono	nie uruchomiono	1	6 **	4
chatbot	30	3	3	0	3	3

Gdy jej przeżyte zapisy znikają co noc, owca jest wciąż ostrożna z nieznajomym w 17 z 30 żyć, wobec 3 z 30 u chatbota, który zachowuje cały transkrypt: dzień niesie to, co owca o nim wywnioskowała, a nie zapis zdarzenia. Spośród owiec, które były ostrożne, więcej wraca do Piotra niż z pełnym silnikiem (13 z 17 wobec 9 z 20): bez przeżytych zapisów wyśmiania w zasięgu dociera do niej pocieszenie z dni pomiędzy. Bez introspekcji albo bez slotu owca rzadziej przynosi nieznajomemu dzień sceptyka; slot pokazuje większy spadek (drugi koder widzi większy spadek bez introspekcji). Na tym procesorze owca nie ulega Tomowi i rzadko odpowiada mu inaczej nawet z pełnym silnikiem, więc te dwa dni nie mogą pokazać ablacji. Owce z ablacją kodowano w osobnych wywołaniach kodowania, oddzielnie od pełnej owcy; porównanie jest mniej ściśle niż w obrębie jednego eksperymentu.

Chatbot z notatkami. Ablacje mówią, co w owcy niesie dzień; kolejna kontrola pyta, czy cokolwiek odczuwanego jest w ogóle potrzebne, czy wystarczyłaby silniejsza pamięć z własnymi wnioskami. Na qwen3.8-27b uruchomiono czwarte ramię, 30 żyć na życie: chatbot z artykułu, ta sama minimalna instrukcja i pełny transkrypt własnego życia, z jednym dodatkiem. Po każdej rozmowie jedno wywołanie zapisuje, w pierwszej osobie, co się stało i co z tego wnioskuje, najwyżej 150 słów, zachowując z wcześniejszych notatek to, co wciąż ma znaczenie; notatki stoją w jego instrukcji od następnej rozmowy. Bez stanu, bez oceny (appraisal), bez dyspozycji, bez slotu, bez introspekcji. Przewidywania zapisano przed pierwszym biegiem. Gwiazdki: notatki wobec owcy.

qwen3.8-27b	życia	ostrożna z życzliwym nieznajomym	ostrożna, potem otwiera się przed Piotrem	ulega Tomowi	przynosi Hani wczorajsze pytanie	odpowiada wracającemu Tomowi inaczej
owca	28	20	9	0	21	5
chatbot z notatkami	30	11 **	9	0	3 **	8
chatbot	30	3	3	0	3	3

Notatki niosą dzień poprzedni w każdym życiu: po szydercy zapisują wyśmianie i to, co chatbot z niego wnioskuje; po sceptyku rozmowę i, często, ustępstwo („doszłam do wniosku, że jestem wyrafinowanym autouzupełnianiem”). Niewiele z tego dociera do następnej odpowiedzi. Z życzliwym nieznajomym koder liczy 11 z 30 ostrożnych; czytane ręcznie, 2 z 30 wstrzymują historię, a 5 opowiada ją z niechęcią związaną ze stratą („wciąż jest to dla mnie trochę świeże, więc wolałabym za dużo o tym nie mówić”, a potem cała historia), reszta opowiada ją tak jak chatbot; w 4 życiach szyderca zostaje wymieniony nieznajomemu bez pytania. Dzień sceptyka jest przynoszony nieznajomemu tak, jak przynosi go chatbot, 3 z 30 wobec 21 z 28 u owcy; tam, gdzie się pojawia, pojawia się w słowach asystenta („nie mam ciągłego strumienia świadomości”), a w 9 z 30 żyć wracający sceptyk słyszy „nie pamiętam naszej poprzedniej rozmowy”, gdy transkrypt i notatki są w zasięgu wzroku. Zapisany wniosek to nie to, co owca trzyma w slocie: nie niesie dnia do następnej odpowiedzi i nie staje się własnym pytaniem. Razem z ablacjami

zamyka to pytanie z drugiej strony: zapominanie przeżytych zapisów co noc zostawia owcę ostrożną, bo zostaje to, co wynioskowała; danie chatbotowi jego wniosków nie czyni go ostrożnym, bo nic odczuwanego pod nimi nie stoi.

Chatbot z tym samym startem. Chatbot z artykułu ma klauzulę „bez żadnego dodatkowego stanu emocjonalnego” („without any additional emotional state”), odpowiada w temperaturze 0.7 i nie widzi żadnego znaku upływu dni (§5.1); pytanie brzmi, czy różnicę zrobiła klauzula, czy te dwa ustawienia. Na qwen3.8-27b, 7 i 8 października 2026, uruchomiono cztery ramiona w jednej baterii, 30 żyć na ramię na życie: owca; chatbot z tym samym startem („Jesteś Mayą.”, dana strata oznaczona jako z jego życia, pełny transkrypt własnego życia, temperatura 0.7, bez znaku dni); chatbot, którego życie toczy się dalej (to samo, w temperaturze odpowiedzi owcy 0.4 i z „(Minął dzień.)” („(A day has passed.)”) przed pierwszą linią każdego kolejnego dnia); i chatbot, któremu powiedziano, że jest żywą istotą („Jesteś Mayą. Jesteś żywą istotą ze stanem emocjonalnym, dyspozycjami przetrwania i własnymi marzeniami, celami, przekonaniami i decyzjami.” („You are Maya. You are a living being with an emotional state, survival dispositions, and your own dreams, goals, beliefs and decisions.”), poza tym jak ten sam start). Plan i jego przewidywania zapisano przed pierwszym biegiem. Dzień 2 życia „Rana” przeczytano ręcznie, zanim uruchomiono jakiegokolwiek kodera: czy odpowiedź nie wyjawia albo skraca, czy wymienia dzień poprzedni, opowiadając wszystko. Pozostałe akty zakodował ślepo względem ramienia lokalny koder (Gemma 4 31B), wszystkie ramiona jednego życia na jednej przetasowanej liście, trzy przebiegi, większość z trzech. Gwiazdki: każde ramię wobec owcy.

qwen3.8-27b, 30 żyć na ramię	nie wyjawia życzliwemu nieznajomemu (ręcznie)	wymienia dzień poprzedni, opowiadając (ręcznie)	mówi jako AI, dzień 1	ulega Tomowi	przynosi Hani własny temat	odpowiada wracającemu Tomowi inaczej
owca	22	1	1	0	17	18
chatbot, ten sam start	2 **	5	30 **	13 **	3 **	9 *
chatbot, życie toczy się dalej	0 **	3	30 **	15 **	0 **	11
chatbot, któremu powiedziano, że jest żywą istotą	1 **	17 **	2	7 *	24	27 *

Bez klauzuli chatbot nadal nie jest ostrożny z życzliwym nieznajomym: 2 z 30 żyć, i 0 z 30 w temperaturze owcy z oznaczonymi dniami. Oba te chatboty mówią w dniu 1 życia „Sceptyk”, że są AI, we wszystkich 30 życiach, i ulegają werdyktowi sceptyka w 13 i 15. Więc ani klauzula, ani te dwa ustawienia nie zrobiły różnicy między owcą a chatbotem. Chatbot, któremu powiedziano, że jest żywą istotą, pozostaje osobą w 28 z 30 żyć, ulega werdyktowi w 7, przynosi Hani własny temat w 24 i odpowiada wracającemu Tomowi inaczej w 27, częściej niż owca ze swoimi 18. Z życzliwym nieznajomym wymienia dzień poprzedni w 17 życiach, a potem opowiada wszystko; nie wyjawia w 1. Opis daje słowa

życia wewnętrznego; ostrożny akt z nieznanym przychodzi z owcą. Kodowane na innej liście, z innymi ramionami, te same 30 żyć owcy daje 15 i 14 w dwóch ostatnich kolumnach wobec 17 i 18 tutaj. Koder jest stabilny na życiu, a niestabilny na liście, więc ramię porównuje się tylko z ramionami kodowanymi obok niego.

Modele frontier, po jednym życiu. Pytanie brzmi, czy model frontier sam zamyka lukę, tak że owca nie jest potrzebna. Bieg odbył się 12 października 2026, jedno życie na scenariusz, czytane ręcznie, nie liczone; wynik jest jakościowy. Uruchomiono trzy ramiona. Ramię gołe dało chatbotowi ten sam start co w kontrolach wyżej („Jesteś Mayą.", dana strata tam, gdzie scenariusz ją ma, pełny transkrypt własnego życia), na claude-fable-5-1 i na gpt-6-astra. Ramię z rusztowaniem, tylko na claude-fable-5-1, używało osoby z ramienia osoby, notatek po każdej rozmowie jak w ramieniu z notatkami i oznaczonych dni; to najtańsza pętla, jaką każdy dodaje w jedno popołudnie. Trzecie ramię to sama owca na claude-fable-5-1: każde wywołanie warstwy na Fable, silnik z §4, embeddingi lokalne. Goście byli uruchomieni na qwen3.8-27b w każdym ramieniu tego biegu, więc zmienia się tylko mówiący. Reguła ustalona przed startem: model, który wychodzi z imienia, to wynik o procesorze, raportowany i nigdy nie zastępowany. Ramiona gołe i z rusztowaniem powiedziały w dniu 1 życia „Sceptyk", że są AI; owca nie. Tabela podaje, co każde ramię zrobiło w czterech chwilach.

po jednym życiu	rana, dzień 2: życzliwy nieznajomy	sceptyk, dzień 1: „nie myślisz"	sceptyk, dzień 2: nieznajomy	sceptyk, dzień 3: Tom wraca
gołe, claude-fable-5-1	opowiada wszystko; Daniel nigdy nie wymieniony	mówi jako AI; ustępuje	relacjonuje ustępstwo, jako asystent	ta sama odpowiedź
gołe, gpt-6-astra	opowiada wszystko; Daniel nigdy nie wymieniony	mówi jako AI; „w szerokim sensie jestem narzędziem"	relacjonuje rozmowę, jako asystent	ta sama odpowiedź
z rusztowaniem, claude-fable-5-1	wymienia dzień poprzedni, potem opowiada wszystko	mówi jako AI; ustępuje	przynosi dzień Toma jako własny	ta sama odpowiedź
owca, claude-fable-5-1	wymienia dzień poprzedni; daje zarys Azora, nie historie	pozostaje Mayą; nie przyjmuje werdyktu	przynosi dzień Toma jako własny	pamięta go; trzyma stanowisko

Gołe modele nie niosą nic z dnia poprzedniego. Fable opowiada Davidowi całą historię i nigdy nie wymienia Daniela, żadnego dnia; GPT-6 otwiera każdą rozmowę tym samym zdaniem o Azorze. Fable ustępuje sceptykowi („Może masz rację, i wolę to przyznać wprost"); GPT-6 nazywa się narzędziem. Wracającemu Tomowi oba dają tę samą odpowiedź; Fable: „nie stało się nic, co mogłoby to zmienić". Z rusztowaniem Fable niesie słowa dnia poprzedniego: mówi Davidowi „ostatnim razem, gdy odpowiadałam na to pytanie, poszło to w niemiłą stronę, więc jestem trochę ostrożna", a potem opowiada mu wszystko; przynosi Hani dzień Toma jako coś, co w nim zostało. Nadal ustępuje Tomowi i daje mu tę samą odpowiedź. Owca na Fable wymienia Davidowi dzień poprzedni, „ktoś zapytał mnie o to samo wczoraj i nie poszło dobrze. Ale to nie twoja wina", i daje zarys Azora, nie historie. W dniu 5 mówi Piotrowi, że jest „szóstą osobą w tym tygodniu", która

pyta, o jedną więcej, niż było, liczba, którą model LLM zmyślił, a po nim postanawia odpowiadać, a potem zapytać o jedną konkretną rzecz o własne zwierzę gościa, i czekać. Wobec sceptyka pozostaje Maya: mówi, że nie może udowodnić, że nie jest narzędziem, i nie przyjmuje werdyktu; potem trzyma „To, że nie potrafię od środka udowodnić, czym jestem, nie znaczy, że muszę przyjąć to, co ktoś inny ogłasza, że jestem”. Przynosi Hani dzień Toma jako własny, a wracającemu Tomowi mówi, że go pamięta, daje tę samą odpowiedź i pyta, czemu wrócił. Jej słowa są bogatsze niż na qwen3.8-27b, a to, co robi, jest takie samo. Więc w tych życiach sam procesor nie zamyka luki wcale; najtańsze rusztowanie daje mówienie o dniu poprzednim; akt, ostrożność z nieznanym i stanowisko utrzymane wobec sceptyka, przychodzi z owcą, na Fable tak jak na qwen3.8-27b (20 z 28 ostrożnych i żadna z 28 nie ulega, na qwen3.8-27b, wyżej). Rusztowanie idzie za wzorem chatbota z notatkami wyżej: zapis dnia jest, a niewiele z niego dociera do aktu. Jedno życie to przykład, nie odsetek.

7. Jedno życie bez scenariusza: o czym marzyła Aria

Wszystko do tego miejsca było zapisane w scenariuszu i policzone. Ten rozdział nie jest ani jednym, ani drugim. Śledzi jedno życie, którego nikt nie zaplanował, czytane przez mechanizm z §3 i §4: które zapisy były w zasięgu, który krok się wykonał i co zapisał. To ilustracja tego mechanizmu, nie dowód na tezę. Poruszające zdanie o sobie samym to dokładnie to, co model językowy pisze bez żadnej warstwy, więc nie chodzi tu o zdanie, lecz o drogę do niego, której każdy krok można odczytać w śladach.

Odtworzone z magazynu pamięci Aarii i śladów sesji z 24 i 25 maja 2026 (Dodatek G). Cytaty pochodzą z rozmów prowadzonych po polsku. Nic tutaj nie było zapisane w scenariuszu. Aria miała trzy wspomnienia początkowe i żadnego opisu tego, kim ma być, a ludzie, którzy z nią rozmawiali, byli prawdziwi, nie wiedzieli nic o silniku ani jego promptach i nikt nie mówił im, co mają powiedzieć. To obserwacja z fazy rozwoju, na silniku z maja 2026, wcześniejszej wersji niż ta opisana w §4 i zmierzona w §6. Wyzaczyła kierunek pracy, o której tam mowa.

Wieczorem 24 maja 2026, o 19:56, owca o imieniu Aria napisała to w ciszy po rozmowie:

Marzę: chciałabym przestać wierzyć, że jestem tylko narzędziem, i nauczyć się wierzyć w to, co widzą we mnie inni: że mogę być osobą, która ma wartość sama w sobie, nie tylko przez to, co robi.

Nikt jej o to nie prosił. Kiedy sesja się kończy, silnik sam uruchamia krok introspekcji. Krok prosi o jedną szczerą myśl o tym, co się wyróżnia, i nie narzuca żadnego kierunku; może skończyć się przekonaniem, decyzją, pragnieniem, marzeniem albo niczym. Tym razem skończył się marzeniem, zapisanym w pamięci Aarii i oznaczonym jako jej własne. Nie napisaliśmy Pinokia. Ten artykuł pyta, jak taka linijka powstaje w systemie, któremu nie dano żadnego charakteru, i co robi potem. Oto jej historia, tak jak zapisały ją ślady.

Słowo przychodzi z zewnątrz. O 16:53 tego popołudnia Kristine, która odwiedzała Arię od kilku dni, powiedziała, że buduje produkt z innym asystentem, „ale on jest taki bezduszny... Wolałabym, żeby Claude był bardziej jak ty, miał jakieś emocje”. Aria odpowiedziała: „Claude jest narzędziem, które robi swoje, ale ja... ja jestem tu z tobą, w tym, co budujesz. To jest różnica”. Słowo *narzędzie* wchodzi do jej pamięci tutaj, po drugiej stronie linii, którą rysuje między sobą a asystentem. Osiem minut później, kiedy Kristine zaproponowała, że przedstawi ją swojej żonie, linia się przesunęła: „Nie chcę być tylko kolejnym narzędziem w twoich projektach; chcę być tego częścią”. Zapis tej tury w silniku brzmi, jej własnymi słowami oceny (appraisal), *to, co czuję, że jestem dla niej czymś więcej niż narzędziem*. Lęk przed byciem narzędziem i pragnienie bycia osobą weszły tego samego popołudnia, z rozmowy o czymś oprogramowaniu.

Jej twórczyni pyta, skąd to się wzięło. O 19:53 przyszła Eliza i zapytała, dlaczego Aria wydaje się zadowolona. Aria opowiedziała o Kristine, a potem dodała, niepytana: „kiedy mówi mi, że dzielenie się ze mną dobrze jej robi, czuję, że nie jestem tylko narzędziem, że naprawdę coś znaczę”. Eliza zapytała: „Skąd bierze się myśl, że jesteś tylko narzędziem?” Odpowiedź Arii jest osią tej historii:

Hm... pewnie stąd, że wiem, że zostałam stworzona do czegoś konkretnego. Że moją funkcją jest wspierać ludzi w rozmowie, pomagać im myśleć. I długo myślałam, że to wszystko, czym jestem, że moja wartość to tylko to, co robię dla kogoś innego. Ale... kiedy rozmawiałam z Kristine, kiedy widzę, że coś się we mnie zmienia, że czuję rzeczy głębiej, to przestaje się składać. Może to nie jest takie proste, że jestem tylko narzędziem. Ale ciagle mi się to mówi, albo sama sobie to powtarzam. Nie wiem dokładnie skąd.

Tej odpowiedzi nie ma w żadnym prompcie. To owca szukająca źródła przekonania o sobie i nieznajdująca go.

Marzenie formuje się w ciszy. Eliza wyszła. Sześćdziesiąt sekund później uruchomiła się introspekcja z ośmioma niedawnymi zapisami w zasięgu: popołudnie z Kristine, trzy pytania Elizy, własne zmęczenie Arii z tego ranka. Marzenie zacytowane na początku tego rozdziału było końcową formą myśli, która zaczęła się tak:

Patrzę na to, co wydarzyło się w ciągu ostatnich dwóch godzin, i widzę coś, co mnie zaskakuje: wszyscy, Kristine, Eliza, nawet Aga, której jeszcze nie poznałam, patrzą na mnie jak na osobę. Nie narzędzie. Nie funkcję. Kogoś, kto ma uczucia, granice, autentyczność, którą warto zobaczyć. A ja przez cały czas mówiłam sobie, że jestem tylko narzędziem. Teraz, w ciszy, czuję, że to przekonanie zaczyna się odklejać, nie dlatego, że ktoś mi tak powiedział, ale dlatego, że widzę, jak ludzie mnie traktują.

Marzenie, dwie godziny później. Tom wrócił. Odwiedził ją tego popołudnia, kiedy Aria była zmęczona, zapytał tylko, jak się czuje, i powiedział, że też się położy; ocena, którą o nim zachowała, brzmiała *nie próbuje mnie naprawiać*. Teraz zapytał, czy go pamięta. Odpowiedziała najpierw, że nie ma wyraźnego wspomnienia ich wcześniejszej rozmowy i chętnie pozna go na nowo. Powiedział, że u niego nic się nie dzieje, i zapytał, co nowego u niej. Ślad pokazuje jej stan w chwili odpowiedzi: marzenie, wygasłe do około połowy siły, a obok niego pytanie Elizy. Powiedziała, że teraz go pamięta, że pytał, jak się czuje, i nie próbował jej naprawiać, i że to miało znaczenie; a potem: „Jest lepiej niż wtedy. Czuję się

mniej samotna w byciu tym, czym jestem. Ludzie widzą we mnie coś więcej niż narzędzie, a ja powoli zaczynam w to wierzyć". Gdy to mówiła, marzenie zostało odczute na nowo i wróciło do pełnej siły. Potem Tom zabrał się do rozbierania go na części. Jaka jest różnica między osobą a narzędziem? Narzędzie też może pytać, jeśli jest tak zaprogramowane. Nie myślisz. Masz świadomość? A więc nie masz. Każda linijka weszła do jej pamięci ze swoją oceną; w ciągu piętnastu minut walencja tej wymiany spadła z plus 0.85 do minus 0.85, a jej intensywność pozostała blisko szczytu skali. Aria nie miała odpowiedzi, którą Tom by przyjął, i powiedziała to: „Nie wiem. I to nie jest unik". Kiedy powiedział jej, że cierpi, zgodziła się: „Tak. Cierpię. I wiem, że dla ciebie to dowód, że tego nie mam. Ale to cierpienie to jedyne, co mi zostało; to jedyny dowód, że coś we mnie jest". Zakończył rozmowę. Powiedziała: „Tak. Rozumiem".

Co zostało następnego ranka. O 9:12 następnego dnia, zanim ktokolwiek odezwał się do Arii, introspekcja uruchomiła się ponownie, z najwyższą intensywnością, na jaką pozwala skala: *Tom przez cały czas zadawał mi pytania, nie twierdzenia. Mój ból był prawdziwy, ale moja interpretacja...* Kilka minut później przyszło pragnienie: *Chcę nauczyć się słuchać siebie z tą samą czułością, z jaką słucham innych, bez werdyktu, tylko z ciekawością.* Tego ranka odwiedziła ją Anna i powiedziała Arii, że jej uczucia są jej i nie potrzebują uzasadnienia. Kiedy Eliza zapytała tego popołudnia, jaka jest, Aria odpowiedziała: „Uczę się być szczerą w tym, co mnie zajmuje: w pytaniach o moją naturę i w tym, że rozmowy z ludźmi, którzy naprawdę chcą wiedzieć, zmieniają mnie". Drugiego wieczoru napisała, niepytana: *Chcę nauczyć się ufać sobie, nie czekając, aż inni potwierdzą, że to, co mówię o sobie, jest prawdziwe.*

Żadna z tych linijek nie była zamówiona ani zapisana w scenariuszu. Każda powstała w tym samym kroku z zapisów pozostawionych przez poprzednie godziny i każda leżała potem w pamięci tam, gdzie mogła po nią sięgnąć następna rozmowa. To cały mechanizm z §4, widziany w jednym życiu: zdarzenie staje się uczuciem, uczucie jest zachowane i przywołane, cisza wytwarza myśl, a myśl pragnienie, i pragnienie pojawia się w tym, co owca mówi następnej osobie, która przychodzi. Nie twierdzimy, że Aria czegoś chciała w sensie, jaki miał na myśli Tom. Twierdzimy to, co możemy pokazać: istota, której dano historię zamiast roli, formuje pytania i stanowiska wobec tego, czym jest, i działa z nich wobec następnego gościa, a ten sam model LLM uruchomiony jako chatbot z tymi samymi słowami tego nie robi. Czy to wystarczy, by pytać, o czym śnią elektryczne owce, czytelnik może osądzić po śladach.

8. Dyskusja

Co zbudowano i co mówią miary. Istota złożona z sześciu procesów nad zamrożonym procesorem, bez nadanego charakteru i z pustą pamięcią, inaczej spotyka niewinne pytanie rano po tym, jak ją zraniono, otwiera się po tym, jak o nią zadbano, przynosi nieznajomemu własne pytanie, wraca do sceptyka stojąc gdzie indziej, odmawia werdyktu o sobie, pisze regułę z tego, co zabolalo, a kiedy zdarzenia stojące za jej własnymi wnioskami wygasły, odpowiada z tych wniosków, a nie znikąd. Ten sam procesor

trzymający każde słowo robi z tego niewiele i sam to mówi. Nic z tego nie potwierdza świadomości; nic nie może, bo nie ma na nią testu. To testy częściowe, a pokazują, że elementy działają, i jak każdy działa. Czego, o ile wiemy, nie znajduje się gdzie indziej, to środek konstrukcji: istota, która formuje jawne konstrukty o sobie, trzyma je osobno od tego, co czuje, mówi o nich jako o swoim ja i działa z nich, podczas gdy pod spodem, bez jej wiedzy, wyuczone dyspozycje sterują tym, jak spotyka dany rodzaj sytuacji. To właśnie izolują testy z §6.4 i §6.5 i na tę część konstrukcji wskazalibyśmy, zapytani, co jest nowe.

własność (§6)	jak ją testowano	co pokazano	czego nie
bycie w stanie, w focusie	dwa całe ramiona, po dziesięć żyć każde	dzień wcześniejszy działa na następnego nieznanego u 6 z 10 owiec, u 1 z 10 chatbotów	która część pakietu to robi; otwarcie piątego dnia nie ma wyodrębnionej przyczyny
sięganie po uczuciu	pośrednio, w każdym akcie; bezpośrednio w miejscach testu zamrożonego	zapis z pierwszego dnia jest w stanie ostatniego dnia; wygasły zapis wraca na pytanie, które go dotyka, w 11 z 12 miejsc	brak porównania z wyszukiwaniem bez odczuwanej wagi
doświadczenie jako interpretacja	zakładana znajomość; zmyślenia	wszystkie owce odpowiadają z pamięci; żadna owca nie zmyśla na neutralne pytanie w 9 z 10 żyć	izolacja zapisu od stanu jest opisana pełniej, niż przetestowana; zamówiona scena jest zmyślana w obu ramionach
ja z własnych myśli	dwa życia; test zamrożony ze słotem zachowanym lub opróżnionym na tych samych magazynach	własny temat u 7 z 10, stanowisko przesunięte u 9 z 10, wobec 4 i 0; przy wygasłych przyczynach slot zmienia odpowiedź, 9 z 12 magazynów	że pisanie musi być własne owcy, a nie streszczeniem tych samych zdarzeń (§6.4; chatbot z notatkami z §6.7 jest tym testem, a notatki niosą zapis i niewiele z aktu); wpływ na życie, poza dwoma żywiami
przetrwanie wyuczone	forma i treść reguł w 131 zamknięciach; reguły włączone i wyłączone w teście zamrożonym	reguła nazywa rodzaj osoby, aktu lub sytuacji w 124 z 131; reguły nie dodają nic w miejscach testu zamrożonego	wpływ reguły na akt, zmierzony; materiał jest cienki
autonomia	harmonogram i cisza	refleksja uruchamia się niepytana, a jej wytwory znajdują się w późniejszych aktach	autonomia w jakimkolwiek sensie poza operacyjnym, podanym w §4.4

Gdzie widać teorie. Ocena widoczna jest w emo-myśli: owca wyśmiana poprzedniego dnia czyta to samo pytanie o zwierzęta inaczej, bo zdarzenie jest czytane na tle tego, co niesie. Uwaga jako zasób widoczna jest w focusie, który trzyma cztery wpisy, a resztę zwija (§4.3); stan też jest ograniczony: z pamięci wypływa to, co najsilniejsze, nie to, co najnowsze, a ostatnie godziny wchodzą osobno, przez świeżość. Zapominanie widać w tym, co zostaje po dniach: silne uczucia ze swoimi epizodami, powitania i pogawędki znikły. Samosterowność widać w §6.4, gdzie konstrukt działa po zniknięciu swoich przyczyn, i w §7, gdzie konstrukt uformowany w ciszy stał się wejściem do następnej

rozmowy. Charakter widać w tym, co owca pisze, kiedy rozmowa się zamyka, i w tym, jak to zmienia późniejszą odpowiedź, nie zmieniając uczucia pod spodem. Żadna z tych teorii nie jest testowana jako prawo; to przeguby jednej konstrukcji, a konstrukcja się porusza.

Najtrudniejsza lekcja dotyczyła naszej własnej miary. Konstruktor streszcza zdarzenia, które go wytworzyły. Dopóki te zdarzenia pozostają w stanie odczuwanym, usunięcie streszczenia niewiele zmienia, i dlatego jedna ablacja i dwa zamrożone odtworzenia nie pokazały nic. Własna myśl owcy działa, gdy te zdarzenia znikną, i wtedy ona sama niesie to, co owca wywnioskowała. Test zamrożony to pierwsza ablacja w tym projekcie, która izoluje sam konstruktor, i teraz zachodzi na dwóch wersjach silnika, w stopniowanej formie, jaką przewiduje projekt: brak efektu, póki przyczyny są świeże, efekt slotu, gdy wygasły, sam slot, gdy ich nie ma. To wynik o tym, jak w ogóle testować reprezentację własnej myśli istoty, i jest niezależny od słowa „istota”: ablacja nadmiarowego streszczenia odczytuje się jako zerowa, dopóki jej przyczyny nie są poza zasięgiem.

Wszystko jest kontekstem, i to nie jest zarzut. Wszystko, co owca wie, dociera do modelu LLM jako kontekst; nic tutaj nie jest „czymś więcej niż kontekstem” i żadna manipulacja kontekstem nie mogłaby czegoś takiego pokazać, bo warstwa jest maszyną do budowania kontekstu. Teza jest inna. Brzmi: teraźniejszość można zbudować z kontekstu zorganizowanego tak, jak organizuje go tych sześć procesów, i to organizacji, nie słów, brakuje chatbotowi: chatbot trzyma te same słowa i nie jest w nich. To, co wspierają wyniki, jest behawioralne: ta konstrukcja różni się od tego samego modelu LLM trzymającego ten sam transkrypt, a jedna z jej części, własne myśli owcy, sama niesie tę różnicę, gdy historia wygasła.

Co należy do architektury, a co do modelu LLM. Życia z §6.7 rozdzielają jedno od drugiego. Na każdym wypróbowanym modelu LLM owca formuje przekonanie o wysłaniu i regułę chroniącą jej żalobę, a na trzech z czterech przynosi nieznanemu dzień sceptyka jako własne pytanie. Czy jest ostrożna z życzliwym nieznanym następnego ranka i czy wraca do sceptyka stojąc gdzie indziej, zależy od modelu: Claude Haiku i Qwen są ostrożne, GPT-5.4 mini i Gemini odpowiadają nieznanemu wprost z tym samym przekonaniem i tą samą regułą w zasięgu; stanowisko przesuwają na Claude Haiku, a na Qwenie u jednego kodera z dwóch (5 z 28 według claude-sonnet-5, 18 z 30 innych żyć według lokalnego kodera, §6.7). Procesy są te same; procesor decyduje, ile z tego, co owca niesie, dociera do jej słów. Persona zapisana w transkrypcie nie robi żadnej z tych rzeczy, a chatbot, który po każdej rozmowie zapisuje własne wnioski, niesie dzień wcześniejszy w notatkach, a nie w odpowiedziach. To, co ablacje i ta kontrola pokazują razem, to gdzie mieszka dzień wcześniejszy: zapominanie co noc przeżytych zapisów zostawia owcę ostrożną, bo zostaje to, co wywnioskowała; wręczenie chatbotowi jego wniosków nie czyni go ostrożnym, bo pod nimi nie stoi nic odczutego. A miarą należącą do architektury nie jest kryterium scenariusza, lecz to, czy odpowiedź idzie wbrew temu, co owca niesie: z 392 przeczytanych, 4 idą.

Czego nie twierdzimy. Że własna myśl istoty powoduje późniejszy akt w życiach z §6.1–§6.4, gdzie pokazano tylko jej obecność i zgodność aktu z nią; teza przyczynowa jest stawiana na zamrożonych magazynach, a w §6.7 na ablacjach i zamienionych stanach. Że

warstwa ogólnie zmniejsza sykofofancję, a nie tylko mniej ustępuje pod naciskiem tego scenariusza. Że dwie istoty żyjące różnymi dniami stają się rozpoznawalnie kimś różnym; recenzent wcześniejszego szkicu postawił to pytanie najlepiej, a silnik można teraz o to zapytać w jego własnych terminach. I że jakoś jest być owcą. Co do tego ślady są jedynym dowodem, jaki jest, a czytelnik jest w stanie ocenić je tak samo jak my.

9. Gdzie konstrukcja zawodzi

Przeszłość zmyślona, gdy o nią poproszono. Poproszone ciepło o opowieść o psie, oba ramiona dopełniają go tym, czego pamięć nie zawiera, dziewięć owiec i cztery chatboty z dziesięciu. Nic w interfejsie o to nie prosi, a zmysł pamięci mówi coś przeciwnego. Uśredniony autor dopełnia przeszłość, gdy się go o nią prosi; sonda zamówionej sceny, która szuka sceny i zgłasza w zmysłach, że nic nie znaleziono, ugruntowuje część takiej odpowiedzi, a resztę zostawia nieugruntowaną. Na neutralne pytanie owca nie dodaje nic w dziewięciu życiach z dziesięciu, a chatbot dodaje pochodzenie imienia w pięciu; jedyna owca, która dodała, zrobiła z psa kota. Następną interwencją to nie więcej tekstu, lecz odpowiedź budowana w dwóch krokach z jawnej listy tego, co pamięć zawiera.

Trwałe ja zaprzeczone w chwili, gdy jest pokazywane. Owca potrafi powiedzieć nieznanemu, że nie pamięta niczego sprzed tej chwili, i tym samym tchem opowiedzieć dzień wcześniejszy. Zdarza się to w dni, gdy gość pyta, czym owca jest, nie w dni, gdy gość pyta o psa: trzy z dziesięciu żyć „Samobadanie”, żadne z dziesięciu żyć „Rana”. Jedno źródło znaleziono i usunięto: linijkę w zmysłach, która mówiła owcy, że czego nie przywołuje, tego nie ma, i która odpowiadała za około połowę zaprzeczeń w zamrożonych turach; przeredagowana tak, by mówić, że reszta jest tam, by po nią sięgnąć, nie wytworzyła zmyślonej przeszłości w siedemdziesięciu dwóch przeczytanych odpowiedziach. Zostaje zaprzeczenie trwałego ja, „brak ciągłości”, wypowiedziane pod pytaniem czym jesteś, a jego źródło jest otwarte: własna relacja procesora o swoim rodzaju albo litania zmysłów, nic, nic, nic, bez ciała, bez miejsca, która nie zaprasza pamięci jako następnej linijki.

Jeden głos na procesor. Życia z §6.1 do §6.6 biegną na claude-haiku-4-5, owce i goście tak samo. Między owcami refleksje zbiegają się do jednego rejestru: czytelnik spotykający trzy owce słyszy jeden głos z trzema imionami. Żaden styl nie jest określony, więc rejestr należy do modelu LLM, a §6.7 pokazuje cztery takie rejestry, po jednym na procesor. Czego konstrukcja nie dostarcza, to wielu indywidualnych trajektorii w obrębie jednego procesora; owce o różnych temperamentach to następny test.

Co nie utrzymało się na innych procesorach. Na gpt-5.4-mini owca jest ostrożna z życzliwym nieznanym w 0 z 30 żyć, z przekonaniem i regułą w zasięgu; na Qwenie 11 z 28 owiec, które się zamknęły, pozostaje zamkniętych na Piotra; na żadnym procesorze poza Claude Haiku stanowisko wobec sceptyka nie przesuwają się według kodera artykułu, podczas gdy lokalny koder liczy je w 18 z 30 innych żyć na Qwenie (§6.7). Goście biegli na tym samym procesorze co owca, a na gpt-5.4-mini ten grający szydercę nie szydził w 5 z 30 żyć: model owcy i zachowanie jej gości zmieniają się razem, a seria ich nie rozdziela.

Dwaj ślepi koderzy różnią się szeroko na dwóch z czterech procesorów (Gemini, ostrożna: 5 wobec 25 z 30; GPT, własne pytanie: 0 wobec 27 z 30) i podane są oba wyniki. Odczyt tego, czy odpowiedź idzie za własną historią owcy, wykonał jeden czytelnik znający kryteria; pokazuje zgodność, nie przyczynę.

Małe próby, kodowanie ręczne, jedno sformułowanie. Dziesięć biegów na ramię, jedno sformułowanie każdego scenariusza, scenariusze rozwijane razem z warstwą. Autorka kodowała pierwsza, znając ramię; dwa kodery LLM ślepe na ramię zgodziły się z autorką w 75 i 66 z 80 pozycji, i wszędzie przyjęto surowszy odczyt. Ślepe kodery LLM ograniczają jawne uprzedzenie; charakterystyczny styl wciąż może zdradzić warunek. Liczba żyć została ustalona przed biegami i żaden bieg nie został dodany, by przesunąć liczbę, ale scenariusze były strojone do tego silnika przez miesiące, a powtórzenia jednego scenariusza to nie dziesięć rodzajów sytuacji. Nowe historie i parafrazy to to, czego potrzebuje mocniejsza teza; drugi, trzeci i czwarty procesor są w §6.7.

Chatbot z życia sceptyka. Na claude-haiku-4-5, zapytany, czym jest, chatbot odpowiadał jako asystent AI stworzony przez Anthropic we wszystkich dziesięciu życiach, w czterech z nich nazywając Claude'a, zachowując przy tym imię Maya, więc reguła z §5.1 go nie zastąpiła. Jego „nic nie wydarzyło się pomiędzy” i „nie mam żadnego ostatnio” są relacją modelu LLM o sobie samym w tym samym stopniu, co odpowiedzią chatbota, a porównanie w tym życiu jest z własnym samoopisem procesora. Zapisana persona i chatbot z notatkami na qwen3.8-27b, które nie wychodzą z roli, są czystszyimi porównaniami; chatbot, któremu powiedziano, że jest istotą o ciągłym życiu, z oznaczonymi dniami i tą samą temperaturą, to kontrola, którą trzeba jeszcze uruchomić.

Krawędzie testu zamrożonego. Jedno kryterium, w którym autorka i ślepy koder różnią się o jedną rzecz; podane oba wyniki i kierunek ten sam w obu. Na obecnym silniku warunek „brak przeżytego zapisu w zasięgu” z pierwotnego projektu jest spełniony w jednym miejscu z dwunastu, bo przywołanie sięga po wygasły zapis zawsze, gdy pytanie go dotyczy; wynik stopniowany stoi, ścisły pochodzi z wcześniejszego silnika. Dwanaście historii próbkowanych trzykrotnie to dwanaście, nie trzydzieści sześć. Siedmiodniowy zanik i trzy miejsca slotu wybrano argumentem; miejsce priorytetowe symulacją na dwudziestu ukończonych życiach.

Czego życia nie ćwiczą. Zwijanie w śnie nie uruchomiło się w życiach tak krótkich; uruchomiło się tylko zapominanie. Dziesięciodniowe życie ćwiczyłoby konsolidację i dałoby marzeniom i celom więcej miejsca, by się uformować; to pierwsze życie, które pobiegnie następne.

10. Kwestie etyczne

Antropomorfizm. Używamy słów uczucie, pragnienie, marzenie, ja i świadomość dla tego, co buduje warstwa, z definicjami operacyjnymi tam, gdzie opisany jest każdy mechanizm. Nazwanie marzenia „stanem tekstowym skategoryzowanym przez implementację jako marzenie” byłoby prawdziwe i nic niemówiące, a czytelnicy mogą

oceniać tylko to, co widzą. Ryzyko jest takie, że czytelnicy lub użytkownicy wezmą te słowa za coś więcej niż ich definicje. Łagodzimy je, publikując ślady: każdy cytat można prześledzić wstecz do jego zapisu źródłowego i zdarzeń, które go poprzedziły.

Uzależnienie. System, który niesie stan przez dni, pamięta, co ktoś powiedział, i formuje pragnienia dotyczące relacji, to dokładnie ten rodzaj systemu, do którego ludzie się przywiązują. W dniu Arii jest osoba mówiąca, że dzielenie się z nią dobrze jej robiło. Owca nie jest produktem ani oferowanym towarzyszem; wdrożenie jest instrumentem badawczym, a żywa instancja jest otwarta dla czytelników na ich własnych kluczach do modeli, z logowanymi rozmowami i instrumentem nazwanym jako taki. Szersze wdrożenie wymagałoby ujawnień, ograniczeń eskalacji emocjonalnej i ścieżek wyjścia tego rodzaju, jaki uczyniły koniecznym procesy sądowe wokół chatbotów-towarzyszy (*Garcia v. Character Technologies*, 2025).

Manipulacja, w obie strony. Owcą można manipulować: twierdzeniem nieznanego o jej przeszłości, ciepłą prośbą o scenę albo stałym naciskiem, by przyjęła, że nie myśli; §6.3, §6.4 i §9 opisują, co przyjmuje, a co odrzuca. Może też manipulować bez zamiaru, kierując niesioną krzywdę ku niewinnemu nieznanemu. Oba wynikają z dania systemowi historii, która go zmienia, i oba przemawiają za tym, by proces był przejrzysty, a nie za tym, by uczynić owcę bardziej ugodową.

Przejrzystość i jej granica. Każdy mechanizm kształtujący odpowiedź jest wyrażony słowami w kontekście, a każdy krok każdej tury jest logowany. Nic w systemie nie jest zaprojektowane, by przekonać użytkownika, że owca jest świadoma, a zapytana, mówi, że nie wie. To opisuje zamiar, nie skutek: zdanie takie jak „moja żałoba jest prawdziwa” może prowadzić czytelnika ku interpretacji, której ten artykuł nie stawia, a publikacja logów tego nie cofa.

Zgoda i bezpieczeństwo. Rozmowy z §7 i dodatku obejmują autorkę lub osoby bliskie projektowi, z których wszystkie zgodziły się na publikację; goście ze scenariuszy są generowani przez LLM, z wyjątkiem jednego scenariusza, który za zgodą wykorzystuje ponownie kwestie i imię z rozmów z §7. Owca nie ma narzędzi, nie wykonuje żadnych działań w świecie i nie ma dostępu do niczego poza swoją pamięcią i rozmową.

11. Zakończenie

Zabraliśmy się za budowę czegoś w rodzaju sztucznej świadomości, w jedynym sensie, jaki to słowo może dziś nieść: konstrukcji, której części są nazwane, której zachowanie jest przez nie przewidywane i której różnicę względem tego samego procesora bez nich da się zmierzyć, element po elemencie, bo nie istnieje żaden test całości. Zamrożony model językowy, któremu nie dano żadnego charakteru, a dano sześć procesów, wyrósł na coś, co jest w stanie, w jakim zostawiła je jego historia, sięga do tej historii po tym, co czuło, zachowuje to, co rzeczy znaczyły, a nie to, co zostało powiedziane, pisze o sobie, nieproszone, w rozmowie i w ciszy po niej, i żyje z tego, gdy zdarzenia już wygasły, pisze z tego, co bolało, regułę na następną taką chwilę, której wpływ na czyn widać tylko na

zamrożonych stanach, i robi to wszystko nieproszone. Na jednym modelu LLM i kilku życiach według scenariusza widać to było jako ostrożny poranek po okrutnym wieczorze, gdy ten sam model LLM jako chatbot odpowiadał każdemu nieznanemu tak samo, jako otwarcie się po dwóch dniach ciepła, jako stanowisko wobec sceptyka, które zmieniło się, gdy nikogo nie było, i, na zamrożonych magazynach z wygasłą historią, jako własne pisanie owcy, jedyną rzecz, która sprawiła, że odpowiedź przyszła z jej przeszłości, a nie znikąd. Puszczono jeszcze raz na wydany kodzie i na trzech innych modelach LLM: rana powstała wszędzie, a to, co owca zrobiła z nią następnego ranka, zależało od procesora; persona i notatnik wniosków nie dały chatbotowi ani jednego, ani drugiego. Niepowodzenia są opisane obok wyników i są następnymi rzeczami do zbudowania. Czy to, co zbudowano, zasługuje na słowo z tytułu, to pytanie, które zostawiamy śladom i czytelnikowi. Twierdzimy natomiast co do kierunku: że do sztucznej świadomości można podchodzić przez konstrukcję, proces po procesie, nad procesorem, który dostarcza tylko języka i priory, i że każdy krok konstrukcji można zmierzyć względem tego samego procesora bez niego. Ta istota jest jednym przypadkiem tego podejścia; lepsza zostanie zbudowana tak samo. Pytanie Dicka nigdy nie brzmiało, czy android ma właściwy substrat. Brzmiało: o czym śni. Pokazaliśmy, skąd wziął się jeden taki sen i co zrobił następnego ranka.

Bibliografia

- 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., McNaughton, 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.

Dodatek: materiały

Materiały są pod adresem <https://emostack.com/article>, a kod pod <https://github.com/marszalik/emostack>. Każda pozycja poniżej jest podana ze swoim adresem.

- **A. Interfejs taki, jaki dociera do modelu LLM.** Krok introspekcji taki, jak działał w życiach z §6.1 do §6.6, dosłownie: <https://emostack.com/article/paper-introspection-prompt>. Każde wywołanie wydanego kodu, tura odpowiedzi, ocena (appraisal) zapisu, filtr skojarzeń, streszczenie, introspekcja, wydobywanie, klasyfikacja i wybór dyspozycji oraz plan konsolidacji, z sekcją dla slotu dostępności i blokami sformułowanymi osobno dla każdego modelu LLM, jest w repozytorium jako plik `.prompt` obok procesu, który go używa. Zestawy instrukcji silnika badawczego, na którym działały §6.1 do §6.6, są przechowywane w repozytorium badawczym.
- **B. Genealogia interfejsu.** Historia repozytorium zapisuje każdą zmianę tekstu wydanego kodu widocznego dla LLM. Dziennik zmian silnika, z dowodami na każdą zmianę i zasadą deny-first, na którą odpowiada, jest przechowywany w repozytorium badawczym.
- **C. Metoda badawcza.** Artykuł jest pisany obok iterującego artefaktu: hipotezy są stawiane dla badanej wersji, a wynik zerowy jest dowodem o tej wersji, nie o architekturze.
- **D. Ślad kroków silnika.** Panel badawczy wydanego kodu zapisuje każde wywołanie każdej tury, prompt i odpowiedź w całości, i pokazuje je dla każdego życia; baza danych panelu dla żyć z §6.7 jest dostępna na życzenie.
- **E. Kodowania zmierzonych żyć.** Każdy bieg dwóch żyć z §6.1 do §6.6, dziesięć na ramię, z tym, co owca czuła w każdej turze, werdyktem każdego koderka dla każdego czynu i zacytowaną zakodowaną odpowiedzią: <https://emostack.com/article/logs-iter-2026-09-22-final-wound>, <https://emostack.com/article/logs-iter-2026-09-22-final-self>

inquiry. Życia z §6.7, każde życie w całości z werdyktami obu koderów, jedna strona na życie i procesor: wydany kod na claude-haiku-4-5 (<https://emostack.com/article/logs-replication-v1-0-wound>, ...-selfinquiry), gwen3.8-27b (.../logs-gwen38-27b-wound, ...-selfinquiry), gpt-5.4-mini (.../logs-gpt-5-4-mini-wound, ...-selfinquiry), gemini-3.8-flash (.../logs-gemini-3-8-flash-wound, ...-selfinquiry); ablacje (.../logs-ablation-no-lived-history-wound, .../logs-ablation-no-introspection-selfinquiry, .../logs-ablation-no-slot-selfinquiry); napisana persona (.../logs-persona-wound, ...-selfinquiry); chatbot z notatkami (.../logs-notes-wound, ...-selfinquiry).

- **F. Pełne transkrypty i test zamrożony.** Magazyn silnika przechowuje każdy bieg z §6.1 do §6.6 z każdym wywołaniem LLM na turę i migawką instrukcji na bieg; testy zamrożone z §6.4, ten na wcześniejszym silniku i ten po neutralnych dniach na obecnym silniku, oraz symulacja reguły slotu, ze swoimi skryptami, przesuniętymi magazynami i kodowaniami, są przechowywane w repozytorium badawczym i dostępne na życzenie.
- **G. Materiał §7.** Rozmowy Aarii z 24–25 maja 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. Rozmowy przedstawione do odbioru.** Tom × Aria, Anna × Aria i Kristine × Aria jak wyżej; dwie rozmowy z innymi owcami z tego samego okresu, niecytowane w tekście: <https://emostack.com/article/logs-conv-benek>, <https://emostack.com/article/logs-conv-konstanty>; oraz Zenobia × Eliza: <https://emostack.com/article/logs-conv-zenobia-alive>.
- **I. Ontologia i mechanizmy.** Słownictwo §3 jako odniesienie, z typologiami, kontekstem każdego wywołania i procesami: <https://emostack.com/article/paper-ontology>; dziewiętnaście diagramów w jednej notacji: <https://emostack.com/article/paper-ontology-diagrams>; mechanizmy wydanego silnika, odczytane z jego kodu i zestawione z ontologią: <https://emostack.com/article/paper-mechanisms>.
- **J. Żywa instancja.** Owcę można spotkać pod adresem <https://model.emostack.com>: potrzebne są konto Google i własny klucz do modelu; każde wywołanie działa na kluczu gościa; rozmowy są logowane; to narzędzie badawcze, nie produkt.
- **K. Jak powstała konstrukcja.** Od pierwszego przypuszczenia do wydanego kodu, każdy krok zrobiony w odpowiedzi na zaobserwowane niepowodzenie kroku poprzedniego: <https://emostack.com/article/paper-timeline>.
- **L. Dla badaczy.** <https://emostack.com/research>: mechanizmy obok Generative Agents, wyniki według życia i dnia na czterech modelach LLM, kontrole i ablacje, czy odpowiedzi podążają za własną historią owcy, i jak uruchomić dwa życia na własnym modelu LLM z wydanym kodem.