REVERIE: AN INFERENCE-TIME COGNITIVE ARCHITECTURE FOR CONTINUOUS LLM DELIBERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs are reactive: a prompt arrives, the model computes, and cognition halts until the next prompt. Existing multi-agent LLM systems inherit this constraint: they coordinate multiple calls per response but still respond only when prompted. We introduce **Reverie**, an inference-time architecture in which specialised agents run continuously on a shared workspace where ideas accumulate, decay, and compete for influence. An executive composes a response when the user speaks, or on its own when accumulated internal activity crosses a threshold; between turns, the system keeps thinking. On a downstream task of research assistance over 60 open-ended scientific questions, five independent human raters, blinded to system identity, preferred Reverie's conversations over Mixture-of-Agents on $201/300$ pairs (67%; mixed-effects 95% CI 0.62–0.72, $p = 1.7\times10^{-7}$), with every rater preferring Reverie. At matched compute, Reverie also produces 16% more conceptually distinct thoughts per turn than Mixture-of-Agents and 106% more than Self-Consistency (Vendi Score, Bonferroni-significant), at roughly $3\times$ lower cloud-equivalent cost per session than Mixture-of-Agents. On nine verifiable benchmarks (factual QA, math, code) Reverie is statistically tied with Mixture-of-Agents on eight and worse on rare-fact retrieval (PopQA), so the cost of continuous deliberation on convergent tasks is small. Although individual thoughts are fragmentary, Reverie's responses match MoA and beat Self-Consistency on naturalness, suggesting that diversity is a property of the deliberative substrate and naturalness a property of synthesis.

1 INTRODUCTION

Multi-agent LLM systems compose several language-model calls per response (Wu et al., 2024; Hong et al., 2023; Wang et al., 2025; 2023). They all share a property with the single-call models they build on. The cognition fires only when a user speaks, runs once, and stops. There is no internal activity between turns, no place for a half-formed thought to live, and no mechanism for the system to decide on its own that something is worth saying. As LLM systems are increasingly deployed on tasks that benefit from sustained reasoning (research assistance, scientific brainstorming, long-horizon planning (Valmeekam et al., 2023)), this prompt-response paradigm becomes an architectural constraint.

Human deliberation does not work this way. When a researcher thinks about a problem, drafts of arguments compete in the background; the decision to articulate one or keep waiting emerges from internal dynamics rather than external triggering (Baars, 1988; Dehaene et al., 1998; Smallwood & Schooler, 2015). Cognition that persists between turns has no analogue in multi-agent LLM systems.

We introduce **Reverie** (Figure 1), an inference-time architecture in which eleven specialised agents run concurrently on a shared workspace where thoughts accumulate, decay, and compete for influence. The eleven agents are four cognitive-style thinkers (analytical, creative, skeptical, empathetic), a critic and censor for quality control, a perceiver, an emotion module, an episodic memory (Minsky, 1980), an initiative monitor, and an executive. The executive composes a response when the user speaks, or on its own when accumulated internal activity crosses a threshold; between turns, agents keep thinking. We call this persistent internal activity the *cognitive substrate*. The architecture draws on Minsky's Society of Mind (Minsky, 1986; 1980) and the global-workspace tradition (Baars, 1988; Dehaene et al., 1998), with modern LLMs filling the role classical instantiations could only fill with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

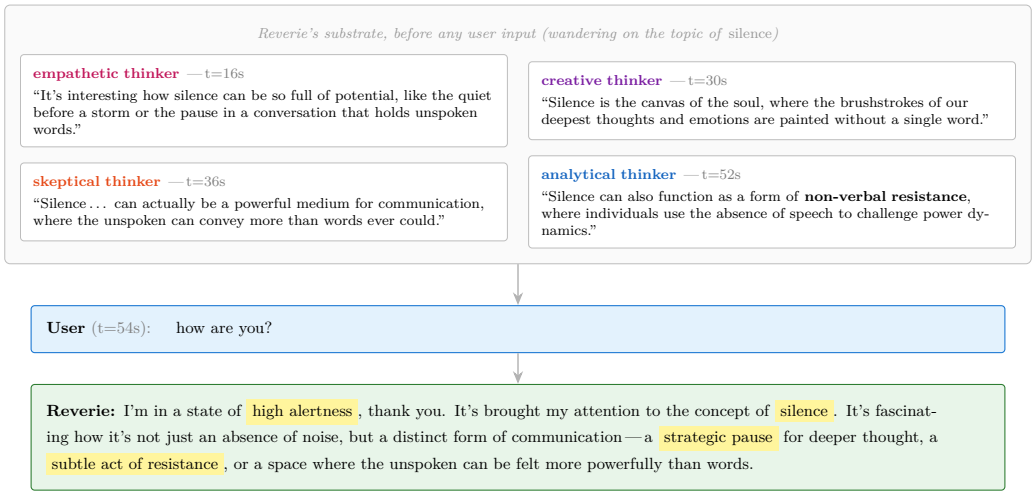


Figure 1: **The cognitive substrate, made visible.** Reverie's four thinkers had been wandering on the topic of *silence* (never raised by the user) for ∼50s when the user said *"how are you?"* at $t = 54\,\text{s}$. The response reports Reverie's emotion (high alertness) and draws fragments from the prior deliberation: strategic pause from the empathetic thinker, subtle act of resistance from the analytical thinker, and the unspoken can be felt more powerfully than words from the skeptical thinker; silence is the shared topic across all four. The response is composed from prior thinking, not from the prompt alone.

brittle hand-crafted knowledge (Lake et al., 2017). Reverie is not a model of human cognition; we claim only that persistent inter-turn deliberation is implementable in a multi-agent LLM system, and that doing so produces measurably preferred behaviour, particularly for open-ended research and for generating a diverse pool of candidate thoughts. Our contributions are:

- We introduce Reverie, the first multi-agent LLM architecture with continuous cognition on a shared workspace and an executive that self-initiates when internal activity crosses a threshold, moving beyond prompt-response interaction.
- To our knowledge, this is the first demonstration that letting a multi-agent LLM system keep thinking between turns yields measurably preferred outputs under blinded human judgment, on open-ended research questions.
- Reverie generates more conceptually distinct thoughts per turn than every multi-call alternative we tested at matched proposer count. To the best of our knowledge, this is the first measurement of substrate-level richness in a multi-agent LLM system.
- Despite Reverie's substrate consisting of fragmentary deliberation rather than complete answer-attempts, its responses match Mixture-of-Agents and clearly beat Self-Consistency on naturalness. Diversity is a property of the substrate, naturalness a property of executive synthesis.

Continuous deliberation helps on open-ended questions, is tied with Mixture-of-Agents on eight of nine verifiable benchmarks at matched compute, and is worse on rare-fact retrieval (PopQA). Reverie is a deliberation architecture, not a replacement for prompt-response systems.

2 Related Work

Cognitive architectures: Classical cognitive architectures (Soar (Laird et al., 1987; Newell, 1990), ACT-R (Anderson et al., 2004), CLARION (Sun, 2005), OpenCog (Goertzel et al., 2014), Minsky's Society of Mind (Minsky, 1986; 1980)) and the global-workspace tradition (Baars, 1988; Dehaene et al., 1998; Franklin et al., 2014) formalised intelligence as the interaction of specialised modules

over a shared workspace. Their structural ideas (modular agents, salience-weighted workspaces, K-line memory) have not, to our knowledge, been re-implemented on a modern LLM substrate, even though the principal limitation of these systems was the brittle hand-crafted knowledge that modern LLMs are now well-suited to supply. Recent interpretability work reports a global-workspace-like structure emerging inside a single LLM, a set of representations the model can report on and reuse across tasks (Anthropic, 2026). Reverie makes the workspace explicit and architectural, shared across agents and persistent across turns.

Multi-agent LLM systems: Multi-agent LLM systems extend single-call models by composing several language-model calls per response. Pipeline frameworks (AutoGen, Wu et al., 2024; MetaGPT, Hong et al., 2023; CAMEL, Li et al., 2023; ChatDev, Qian et al., 2024), aggregation methods like Self-Consistency Wang et al., 2023 samples several reasoning paths from the same model and takes the majority-vote answer, while Mixture-of-Agents Wang et al., 2025) combines multiple LLMs to produce a single superior response, iterative-revision methods (Multi-Agent Debate, Du et al., 2024; Reflexion, Shinn et al., 2023; Self-Refine, Madaan et al., 2023), reasoning-action loops (ReAct, Yao et al., 2023b), and search-style methods (Tree of Thoughts, Yao et al., 2023a) all run once per prompt and stop. None maintains internal state between turns, and none decides on its own that something is worth saying.

Persistent-state agents: Generative Agents (Park et al., 2023) add long-term memory and periodic reflection to a population of simulated characters in a shared environment; MemGPT (Packer et al., 2023) extends LLM context with a hierarchical memory system, and Voyager (Wang et al., 2024) maintains a growing skill library across a lifelong-learning trajectory. Mixture-of-Experts (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022) decomposes computation across specialists but routes inputs through learned gating in a single forward pass. None of these models a single cognitive entity composed of asynchronous sub-agents whose internal dynamics determine when, and whether, to speak. That design point (an individual deliberator sustained across time, not a crowd of personas or a one-shot routing decision) is what Reverie targets.

Our position: Reverie is, to our knowledge, the first multi-agent LLM system in which agents run continuously on a shared workspace and the executive can self-initiate when internal activity crosses a threshold. It combines three commitments that have not previously been combined: the modular-workspace structure of classical cognitive architectures (recently re-examined for LLM agents in Sumers et al., 2024; Andreas, 2022), the LLM substrate of modern multi-agent systems, and the persistent-state insight of Generative Agents. Self-Consistency (SC) and Mixture-of-Agents (MoA) serve as our primary multi-agent baselines (§4); Multi-Agent Debate is excluded because its convergence-driven mechanism is the opposite of the within-turn diversity we measure (Appendix G).

3 The Reverie architecture

Reverie's architecture is summarized in Figure 2. Eleven agents, inspired by Society of Mind (Minsky, 1986; 1980) and global workspace tradition(Baars, 1988; Dehaene et al., 1998) , share a single time-decaying workspace and run on a 2-second tick loop, and the executive composes responses either to user input or to autonomously triggered initiative signals. Full per-agent specifications (prompts, formulas, hyperparameters) are in Appendix I.

Blackboard: The blackboard is a shared workspace storing typed atomic units we call *nemes*. Each neme has content, a salience score $\in [0, 1]$ set at creation, an activation that decays by 2% per tick (half-life $\sim$1.1 minutes), a time-to-live, and metadata. Each neme's effective attention is activation $\times$ salience, and reads return nemes sorted by attention. This gives the architecture soft selection without explicit routing. Garbage collection prunes nemes once activation falls below 0.01 ($\sim$7.5 minutes from creation). A small set of non-decaying fields holds conversation history, valence-arousal, initiative pressure, and active memory IDs.

Per-tick activity: On each tick, the perceiver extracts percepts from any new user input, the four thinkers post thoughts, the critic scores recent thoughts, the censor suppresses noise, memory retrieves K-lines, emotion updates the valence-arousal state, and the initiative agent updates its pressure scalar. Noise here means near-duplicates, low-coherence text, and refusal markers. Different agents have different tick intervals. They run concurrently, not sequentially. The executive is event-

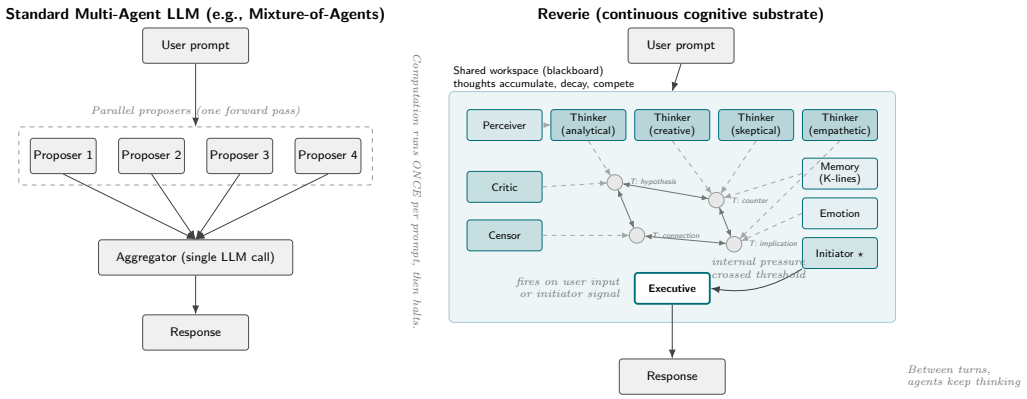


Figure 2: **Reverie (right) versus a standard multi-agent LLM (left).** Standard systems run once per prompt and halt. Reverie's eleven agents run continuously on a shared workspace, with the executive firing on user input or on internal initiative pressure. The curved arrow is the architectural commitment: between turns, agents keep thinking.

driven and runs only when triggered, by user input or the initiative threshold. The thinkers, critic, and executive issue LLM calls, while the perceiver, censor, memory, emotion, and initiative modules are rule-based or computational.

Thinkers: The four thinkers (analytical, creative, skeptical, empathetic) run at staggered tick intervals with temperatures sampled uniformly from $[0.5, 0.9]$. In *active mode* they respond to fresh percepts, and in *wandering mode*, when no fresh percepts are present, they generate autonomous thoughts conditioned on emotional state and recent topics. Wandering thoughts have lower salience and shorter time-to-live vs. active thoughts, but are the raw material from which autonomous initiative emerges between turns.

Emotion, memory, initiative: The emotion module maintains a Russell circumplex valence-arousal vector (Russell, 2003) that decays toward baseline and is injected into thinker prompts as a tone modifier. Memory retrieves K-lines (Minsky, 1980) (snapshots of prior cognitive state, retrievable as episodic priming) by a weighted combination of word overlap and valence-arousal proximity, with no LLM call, and after a successful exchange it captures the active agent state, emotion, and conversational context as a new K-line. This gives Reverie episodic priming without gradient updates. The initiative agent computes a scalar pressure from five signals (uncommunicated salient thoughts, elapsed time since the last user message, arousal, active K-lines, uncommunicated high-salience creative thoughts) and fires when pressure exceeds 0.7, with a 60-second cooldown to prevent rapid re-firing.

Executive: The executive is an LLM-based agent (cloud or local) that composes responses. It is event-driven: it activates when a user message arrives or when the initiative agent fires, waits $\sim$40 seconds for thoughts to accumulate, and then reads the top-ranked thoughts, critiques, and memory recalls to synthesize a response. We deliberately withhold the multi-turn conversation history from the executive: each turn is a single Q$\rightarrow$A call, and cross-turn state is carried entirely by the cognitive substrate. This mirrors the call structure of the baselines (§4) and ensures that any Reverie advantage comes from the substrate, not from extra context.

4 Experimental protocol

We run four experiments. The first measures within-turn and cross-turn diversity across multi-turn conversations, comparing Reverie to four baseline architectures and three internal ablations on a fixed topic set. The second measures human preference on open-ended scientific questions, comparing Reverie head-to-head with a matched Mixture-of-Agents baseline under blinded judgment by five external raters. The third tests the same system where correctness is scored (§6.6), and the fourth sweeps its hyperparameters and role set (§6.7).

Conditions: All experiments share a `gemini-2.5-flash` (Google DeepMind, 2025) executive backbone so that architectural effects are isolated from model capability. We compare Reverie against two single-call baselines (B1 Gemini-Chain-of-Thought (Gemini-CoT) (Wei et al., 2022) and B2 Gemini-Thinking, with extended reasoning enabled) and two multi-agent baselines: *Mixture-of-Agents* (MoA) (Wang et al., 2025), with 4 proposers at $t = 0.7$ (configuration in Appendix L), and *Self-Consistency* (SC) (Wang et al., 2023), with 4 Qwen2.5:32b (Qwen Team, 2024) proposers at temperatures $\{0.5, 0.7, 0.7, 1.0\}$ (mean 0.725, matching Reverie's thinker mean). Both multi-agent baselines feed a Gemini-2.5-Flash aggregator, and all proposers are capped at 150 output tokens. Three internal ablations progressively remove components: A1 (memory and emotion), A2 (critic and censor), and A3 BareThinkers (Reverie's four thinkers and executive retained, but the substrate stripped entirely, the cleanest substrate-vs-no-substrate comparison). Full configurations are in Appendix H.

Conversation history: Single-call baselines (B1, B2) receive the full multi-turn history at each turn, following standard chat-model practice. Reverie, MoA, and SC receive only the current turn on the answer side, and for Reverie this is intentional, since cross-turn state is carried by the substrate (§3). The simulated user always sees the full prior conversation, identically across conditions.

Experiment 1: multi-turn diversity. We use 40 fixed seed topics across four domains (philosophy, scientific reasoning, practical, and creative, with 10 each), and each topic is run for 12 turns under each condition (below). Follow-up questions are produced by a simulated user (`gemini-2.5-flash`, $t = 0.7$, the standard temperature for open-ended generation (Holtzman et al., 2020)), identical across conditions. Because each follow-up depends on the assistant's prior response, the question streams diverge across conditions, and we therefore evaluate whole conversations rather than per-turn matched pairs.

Experiment 2: human preference, 60 questions $\times$ 5 raters. Each of 60 open-ended questions is run as a 12-turn conversation with the simulated user under Reverie and under MoA at a matched per-turn budget (MoA configuration in Appendix L and study protocol in Appendix J). MoA is the comparator because it is the strongest baseline in Experiment 1 (closest on Vendi, tied on naturalness). Each conversation is condensed to a $\leq$200-word summary by `gemini-2.5-flash`. Five raters unaffiliated with the project, blinded to system identity and hypothesis, judge the two summaries per question in a forced choice, with order and A/B assignment randomised per rater.

Experiment 3: verifiable tasks. Reverie vs. MoA on nine public benchmarks $\times$ 100 items, single shot, with identical questions and output-format instructions and matched models, token caps and time budget (MoA configuration in Appendix L and setup in Appendix K). Significance is a paired bootstrap over items (10,000 resamples). Datasets: TriviaQA (Joshi et al., 2017), NQ-Open (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023), GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021), SVAMP (Patel et al., 2021), HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), MBPP+ (Liu et al., 2023).

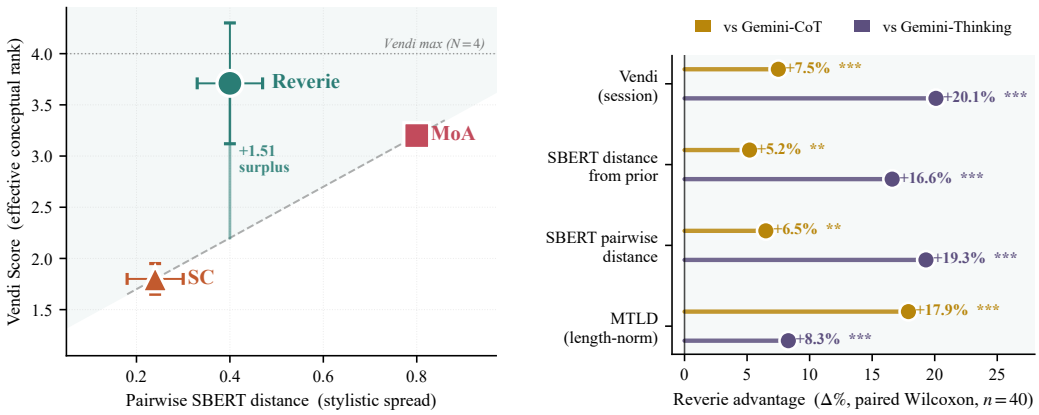
Experiment 4: sensitivity sweep. A smaller version of the Experiment 1 protocol, in which 21 configurations $\times$ 10 topics (not 40) $=$ 210 full 12-turn sessions vary decay rate, executive wait, initiative threshold, cooldown and the role set one at a time, plus a decay$\times$wait grid, scored with the two headline metrics (within-turn Vendi and cross-turn drift).

5 Evaluation framework

We evaluate each system at two levels (thought-level and response-level), and report two families of measurements: diversity, and naturalness under a blinded LLM judge. The downstream human evaluation (Experiment 2, §4) is reported alongside.

Two levels: L_1 (thought-level): the per-turn set of proposer outputs (thought nemes in Reverie and its ablations; proposer outputs in MoA and SC). Undefined for the single-call Gemini baselines. L_2 (response-level): the executive output shown to the user, defined for all conditions.

Diversity: These metrics capture the architectural signature of Reverie's substrate: how richly the agents generate distinct thoughts within each turn, and how well the substrate keeps a conversation anchored across turns. We report three classes. *Lexical*: Distinct-1 and Distinct-2 (Li et al., 2016) measure token unigram and bigram diversity, and MTLD (Measure of Textual Lexical Diversity)



(a) **Reverie produces more conceptually distinct thoughts** (higher Vendi is better). $N = 4$ proposers strictly matched ($n = 40$, MiniLM, ± 1 SD). Dashed line connects the two multi-call baselines (SC, MoA); Reverie's vertical separation ($+1.51$ Vendi) is the dissociation.

(b) **Reverie wins every length-controlled metric** against single-call Gemini baselines (positive Δ is better; paired Wilcoxon, $n = 40$). All 8 contrasts survive Bonferroni correction over the 37-test family. *** $p < 10^{-3}$; ** $p < 10^{-2}$.

Figure 3: **Reverie's diversity advantage at both levels of analysis** (App. A reports full multi-comparison corrections).

(McCarthy & Jarvis, 2010) measures length-normalised lexical richness; together these catch the canonical failure mode of dialogue LLMs (generic, repetitive output), and we treat MTLD as primary within the class because it is length-normalised. *Semantic*: pairwise Sentence-BERT (SBERT)cosine distance (Reimers & Gurevych, 2019) captures average meaning spread within a per-turn set, SBERT-from-prior captures within-conversation novelty (turn t vs turn $t-1$), and cross-turn drift (distance from the turn-1 to the turn-t thought centroid) is the headline anchoring metric, meaningful only for architectures that maintain state between turns. *Entropy-based*: the Vendi Score (Friedman & Dieng, 2023) gives the effective number of distinct items in a set (the exponentiated entropy of its similarity-kernel eigenvalues, $\in [1, N]$) and distinguishes uniformly-spread sets from cluster-spread sets that have high mean pairwise distance but few effective groups (the *Vendi-vs-pairwise dissociation*, §7); we treat it as the primary diversity metric overall. Semantic and entropy metrics are reported under two embedders (`all-MiniLM-L6-v2`, `all-mpnet-base-v2`) for robustness.

Naturalness: We test whether Reverie's executive output reads as human-written despite a fragmentary substrate, using pairwise human-like preference under a blinded Claude Sonnet 4.6 (Anthropic, 2025b) judge with positional, verbosity, and self-preference bias mitigation (Zheng et al., 2023; Bavaresco et al., 2025). Each pair is judged twice with the order swapped. a Reverie win counts only if both orders agree (a "stable win"), giving $n = 40$ pair-level outcomes per comparison.

Statistics: Cross-condition tests use the Mann–Whitney U statistic on per-topic values; paired ablations use Wilcoxon signed-rank; naturalness use a two-sided binomial test against the null of 0.5. Bonferroni correction is applied jointly over the 37-test family reported in §6 (full correction table in Appendix A). All headline cognitive-substrate findings survive Bonferroni correction at $p < 10^{-3}$ to $p < 10^{-12}$. The five-rater study (§6.3) contributes 300 correlated judgments, so its primary test is a mixed-effects logistic model with a rater random intercept, cross-checked with a generalised estimating equation (GEE) with cluster-robust standard errors by rater (Liang & Zeger, 1986).

6 Results

6.1 Reverie vs multi-call baselines

We measure within-turn thought diversity at strictly matched $N = 4$ proposers per turn against the two multi-call baselines (Mixture-of-Agents and Self-Consistency). Figure 3a plots Vendi Score against pairwise SBERT distance for all three conditions ($n = 40$ topics, MiniLM embedder; cross-

Table 1: **Five-rater preference study** (60 questions, forced choice between Reverie and MoA conversation summary, blinded, order and A/B randomised per rater). Raters are anonymised R1–R5.

	R1	R2	R3	R4	R5	Pooled
Reverie preferred	46	44	38	37	36	$201/300$
MoA preferred	14	16	22	23	24	$99/300$
Reverie %	77	73	63	62	60	67
Mixed-effects logistic, rater random intercept: $P(\text{Reverie}) = 0.672$, 95% CrI $[0.617, 0.723]$						
GEE, cluster-robust by rater: 0.670, 95% CI $[0.609, 0.726]$, $p = 1.7\times10^{-7}$						
Within-rater correlation 0.004; majority-of-raters Reverie on $47/60$ items; Fleiss $\kappa = 0.043$						

turn drift is not comparable across these architectures because MoA and SC are stateless across turns, and is reported instead in §6.4 against the substrate-isolating BareThinkers ablation).

Reverie produces effectively more distinct thoughts than both multi-call baselines: $+15.8\%$ Vendi vs MoA ($p = 1.77\times10^{-6}$) and $+106.1\%$ vs SC ($p = 1.44\times10^{-14}$), both Bonferroni-significant.

The Vendi-vs-pairwise dissociation: The figure 3a also exposes a non-obvious pattern: MoA's mean pairwise SBERT distance is twice Reverie's, yet its effective-distinct-items count is lower. The dissociation reflects a stylistic-vs-conceptual diversity distinction. MoA's four proposers span two model families (two Anthropic Claude models; two Google Gemini-Flash-Lite models) and cluster into two stylistic groups: high mean pairwise distance, but low effective rank. Reverie's four proposers come from a single model (Qwen2.5:32b) but are prompted to embody distinct cognitive styles, spreading more uniformly across the embedding manifold. We discuss the implication for downstream synthesis in §7.

The cross-level inversion: Under the blinded Sonnet 4.6 judge with order-swapped stable wins (§5), Reverie is preferred over SC at the response level on $39/40$ pairs ($p = 7.5\times10^{-11}$) and ties MoA on naturalness alone. At the *thought* level, however, both baselines beat Reverie ($0/39$ vs MoA, $0/40$ vs SC): Reverie thinkers produce hedged deliberation fragments while baseline proposers produce complete answer-attempts, and the judge reads completeness as more human-written. The executive synthesis is what closes the response-level gap, weaving distinct fragments into natural output — naturalness is a property of synthesis, not substrate (§7).

6.2 Response-level diversity vs single-call baselines

Single-call baselines have no per-turn proposer set, so only response-level (L_2) diversity is comparable. Reverie wins every length-controlled metric against both single-call Gemini baselines (Figure 3b; paired Wilcoxon, $n = 40$ matched topics, MiniLM; full per-metric breakdown in Table 4, Appendix B). All eight contrasts survive Bonferroni correction. Reverie is shorter than CoT (-18% tokens) and longer than Thinking ($+44\%$); the length-normalised MTLD result and the semantic and distributional metrics remove length as a confound in both directions.

6.3 Human preference: 60 questions, 5 external raters

Experiment 2 (§4) gives 300 blinded judgments (Table 1). All five raters prefer Reverie, at a rate of 60–77%. Because each rater contributes 60 correlated judgments, the primary estimate is the mixed-effects model, and population preference for Reverie is 0.672 (95% credible interval $[0.617, 0.723]$), the GEE cross-check too gives 0.670 (95% CI $[0.609, 0.726]$, $p = 1.7\times10^{-7}$). On a per-item basis, a majority of raters chose Reverie on $47/60$ questions; inter-rater agreement is low (Fleiss $\kappa = 0.043$ (Fleiss, 1971)), so the preference is consistent in direction but noisy on any single item.

6.4 Cross-turn drift and substrate attribution

We isolate the substrate's contribution by comparing Reverie to two internal ablations on cross-turn novelty drift (paired Wilcoxon, $n = 40$ matched topics, MiniLM, ALL-turns window). Removing the substrate entirely (A3 BareThinkers) drives a -31.4% drift reduction ($p = 1.6\times10^{-7}$, Bonferroni-significant); removing memory and emotion alone (A1 NoMemEmo) shrinks the gain

Table 2: **Verifiable tasks, Reverie vs. MoA** (100 items per dataset, single shot, matched models, token caps and 40 s budget). p from a paired bootstrap (10,000 resamples). Calls = mean local 32b LLM calls per item; tokens = mean generated tokens per item, local 32b proposer/thinker plus the shared Gemini-Flash executive/aggregator (prompt tokens excluded). Bold p: significant at 0.05.

		Accuracy			Calls		Gen. tokens	
Task	Dataset	Reverie	MoA	p	Reverie	MoA	Reverie	MoA
Factual QA	TriviaQA	0.76	0.83	0.080	10.4	61.1	701	919
	NQ-Open	0.42	0.43	0.905	10.7	47.2	674	972
	PopQA	0.20	0.33	**0.001**	10.8	47.7	671	979
Math	GSM8K	0.94	0.89	0.055	10.0	7.7	982	1,254
	MATH-500	0.51	0.44	0.206	9.1	7.2	1,283	1,401
	SVAMP	0.89	0.92	0.331	10.0	10.7	827	1,198
Code	HumanEval	0.84	0.87	0.486	9.5	17.4	1,074	1,212
	MBPP	0.76	0.84	0.067	9.7	12.9	914	1,173
	MBPP+	0.80	0.88	0.055	9.8	13.4	929	1,169

by half (-14.2%, $p = 0.028$). The substrate itself, peer-aware blackboard reads, salience-weighted attention, is doing the work. Local diversity does not cost response-level coherence: Reverie and BareThinkers do not differ on embedding-based response-level local coherence or on cosine similarity to the seed topic ($p = 0.19$ and 0.13). Full ablation panel (A2 NoCritCensor) in Appendix E.

Embedder robustness: All embedding-dependent findings replicate on `all-mpnet-base-v2` within $\pm 10\%$ of MiniLM effect sizes (Appendix D). Lexical metrics are embedding-free.

6.5 Cost and scale

Cost: We charge every condition's LLM output at cloud-equivalent per-token rates (\$0.16–\$5/M; full rate sheet and call counts in Appendix C, Table 5). Reverie Pareto-dominates MoA: $+15.8\%$ Vendi at $\sim 3\times$ lower per-session cost (Figure 4 in Appendix C). SC is cheapest in absolute terms but its diversity collapses (Vendi 1.80 vs 3.71). Reverie thinkers run on local Qwen GPUs in our setup; the cloud-equivalent figure is what they would cost via a cloud Qwen endpoint. Wall-clock latency (~ 40 vs ~ 10 sec/turn) is discussed in §8.

Scale: Varying Reverie's thinker model across {3b, 7b, 14b, 32b} with executive held at Qwen2.5:32b (full per-scale means in Appendix C, Table 6): within-turn Vendi saturates at 7b (3b sits 30–36% below the cluster, all three contrasts Bonferroni $p \leq 2.5\times10^{-5}$; 7b, 14b and 32b are not separable after correction, the closest pair being 14b vs 32b at raw $p = 0.02$). The headline result is on cross-turn drift: *none* of the 6 pairwise comparisons reaches significance ($|\Delta| \leq 5\%$, all Bonferroni $p = 1.00$). Drift is scale-invariant — the architectural mechanism, not the per-call thinker capability, is what reduces drift.

6.6 Verifiable tasks: where the advantage stops

Table 2 reports Experiment 3. On eight of nine datasets Reverie and MoA are statistically tied (paired bootstrap); the one significant difference is PopQA, a rare-fact retrieval set, where MoA is better (0.33 vs 0.20, $p = 0.001$). Reverie's higher means on GSM8K and MATH-500 are not significant ($p = 0.055$ and 0.21). Continuous deliberation does not collapse on convergent tasks, but it does not help there either.

The cost side is different. MoA proposes serially within the 40 s budget, so its call count is set by answer length: $\sim$7 calls fit on math ($\sim$140 tokens each), 47–61 on factual QA ($\sim$15–20 tokens each). Reverie issues $\sim$9–11 short thoughts per item regardless of task, and counting the local 32b calls and the shared Gemini-Flash executive/aggregator, MoA generates 1.1–1.5$\times$ more tokens than Reverie on every dataset (local tokens alone: 1.6–1.8$\times$). PopQA is the case where many cheap samples buy rare-fact recall that deliberation cannot. Planning benchmarks were run but dropped (Appendix K).

6.7 SENSITIVITY, AND SUBSTRATE VERSUS PROMPTS

Experiment 4 sweeps 21 configurations on 10 topics each (full results and Table 11 in Appendix K). Only two parameters move within-turn Vendi, and both control how many thoughts accumulate before the executive reads: faster decay lowers it (0.10: 2.51 vs the default 3.24, $p = 0.0006$) and a shorter executive wait lowers it (20 s: 2.41, $p = 0.0002$); together they span 1.91 to 3.66 in the decay×wait grid. Initiative threshold and cooldown move neither metric (all $p \geq 0.38$), and cross-turn drift stays within 0.28–0.37 across every configuration with no significant contrast. Roles matter: dropping the creative or the skeptical thinker lowers Vendi by 13–14% ($p = 0.021$ and 0.011), while swapping roles has no effect ($p = 0.73$). The roles are not the whole story, however. The BareThinkers ablation of §6.4 is Reverie with no substrate (the same four thinker prompts, model, token cap and executive; no blackboard, tick loop, memory, emotion, critic or censor), posting as many thoughts per turn as Reverie on the 40 Experiment 1 topics. Reverie exceeds it by +22.5% within-turn Vendi over all turns and by +44.7% over turns 1–4 (MiniLM), so the diversity advantage is substrate-driven, not a product of the persona prompts alone.

7 DISCUSSION

Five blinded, unaffiliated raters prefer Reverie over MoA two to one on overall quality (§6.3), a different question from the naturalness tie of §6.1, and the intrinsic results suggest why. Reverie's substrate carries more conceptually distinct thoughts per turn, the gain survives with the persona prompts held fixed (§6.7), and an aggregator extracts more from four distinct ideas than from four phrasings of two. This is also why MoA wins on pairwise distance while Reverie wins on Vendi: the two metrics measure stylistic and conceptual heterogeneity respectively. Naturalness, by contrast, is a property of synthesis rather than substrate: Reverie's hedged thought fragments read less naturally than a complete proposer answer, yet the executive closes the gap at the response level, and the judge measures stylistic completion, not the cognition behind it (Bavaresco et al., 2025). Cross-turn drift is flat from 3b to 32b thinkers (§6.5) and across every sweep configuration (§6.7), so anchoring reduction comes from how deliberation is organised across turns (continuous thinking, salience-weighted blackboard reads (Baars, 1988; Dehaene et al., 1998), K-line memory (Minsky, 1980)), not from per-call capability; the critic and censor matter only in the autonomous regime of Appendix F. Finally, on tasks with one correct answer Reverie ties MoA on eight of nine datasets and loses on rare-fact retrieval (§6.6): divergent deliberation is the wrong tool when the task rewards convergence, and its advantage is confined to open-ended deliberation, the setting every human-preference result here measures.

8 LIMITATIONS

Reverie's main practical cost is latency: $\sim$ 40 s/turn (the executive wait window). The evaluated configuration is a research instrument, not production-ready. The simulated user causes question streams to diverge across conditions, so we evaluate conversations holistically (except naturalness, evaluated per-message). Coverage spans 40 topics for the diversity experiments, 60 questions for the human study and nine verifiable benchmarks. On PopQA, divergent thinking is a deficit rather than an asset and Reverie is slightly worse on rare-fact recall (§6.6).

9 CONCLUSION

Reverie shows that prompt-response interaction is not an architectural necessity: eleven continuously running specialized agents on a shared workspace, coordinated by an executive triggered by user input or internal initiative pressure, can be implemented at inference time and produce deliberation preferred over a strong matched-cost baseline. The measured mechanism behind the intrinsic results is substrate richness: the shared workspace sustains more conceptually distinct thoughts per turn than all tested alternatives at lower cost. The system, evaluation framework, and per-session traces are released as supplementary material; we hope these enable behavioural studies of LLM cognition beyond the prompt-response paradigm.

AI USE STATEMENT

Language models are the components of the studied system and of the evaluation pipeline (simulated user, LLM judge etc.), as described in §4 and §5. We have not used generative AI tools to generate research ideas, experimental results or human-rater judgments. We used generative AI only for spell-checking. We take responsibility for the final content of this work.

ETHICS STATEMENT

The human study involved adult volunteer raters judging machine-generated text on non-sensitive scientific questions; no personal data was collected, and raters are anonymized. Reverie's inter-turn deliberation supports tasks where divergent thinking matters, scientific brainstorming, creative writing, etc. by surfacing multiple conceptual angles. The cautions are: (i) The autonomous-initiative mechanism lets the system speak unprompted, so deployment requires guards beyond the 0.7 pressure threshold and 60-second cooldown. (ii) Naive scaling carries a non-trivial compute footprint; (iii) it inherits the biases of its underlying LLMs.

REPRODUCIBILITY STATEMENT

Per-agent specifications, formulas and hyperparameters are in Appendices H and I; the rater instructions, rubrics and question lists in Appendix J; and the setup of the verifiable-task and sensitivity experiments (benchmarks, caps, grader, bootstrap, sweep configurations) in Appendix K, and the summary prompts of the preference study in Appendix J. The supplementary material contains the python scripts and the raw traces for experiments.

REFERENCES

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111 4:1036–60, 2004. URL https://api.semanticscholar.org/CorpusID:186640.

Jacob Andreas. Language models as agent models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.423. URL https://aclanthology.org/2022.findings-emnlp.423/.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, https://www.anthropic.com/news/claude-3-family, 2024.

Anthropic. Claude Haiku 4.5. Technical report, https://www.anthropic.com/news/claude-haiku-4-5, 2025a.

Anthropic. Claude Sonnet 4.6 system card. https://www.anthropic.com, 2025b.

Anthropic. A global workspace in language models. Research post, July 2026. https://www.anthropic.com/research/global-workspace, 2026.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp.

238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mo Bavarian, Clemens S. Winter, Phil Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021. URL https://api.semanticscholar.org/CorpusID:235755472.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John D. Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24): 14529–14534, 1998. doi: 10.1073/pnas.95.24.14529. URL https://www.pnas.org/doi/abs/10.1073/pnas.95.24.14529.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Joseph Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76: 378–382, 11 1971. doi: 10.1037/h0031619.

Stan Franklin, Tamas Madl, Sidney D'Mello, and Javier Snaider. Lida: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, 6 (1):19–41, 2014. doi: 10.1109/TAMD.2013.2277589.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1.

Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. *Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI*. Atlantis Press, 2014.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, https://deepmind.google/technologies/gemini/, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021. URL https://api.semanticscholar.org/CorpusID:232134851.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *arXiv e-prints*, art. arXiv:2308.00352, August 2023. doi: 10.48550/arXiv.2308.00352.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

John Laird, Allen Newell, and Paul Rosenbloom. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33:1–, 09 1987. doi: 10.1016/0004-3702(87)90050-6.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. ISSN 00063444. URL http://www.jstor.org/stable/2336267.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Philip M. McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392, 2010. URL https://api.semanticscholar.org/CorpusID:42852342.

Marvin Minsky. K-lines: A theory of memory. *Cognitive Science*, 4(2):117–133, 1980.

Marvin Minsky. *The Society of Mind*. Simon & Schuster, 1986.

Allen Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560, 2023. doi: 10.48550/ARXIV.2310.08560. URL https://doi.org/10.48550/arXiv.2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

Qwen Team. Qwen2.5 technical report, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

James A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110 1:145–72, 2003. URL https://api.semanticscholar.org/CorpusID:2890641.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jonathan Smallwood and Jonathan Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66:487–518, 01 2015. doi: 10.1146/annurev-psych-010814-015331.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=1i6ZCvflQJ. Survey Certification, Featured Certification.

Ron Sun. The clarion cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, pp. 79–100, 12 2005. doi: 10.1017/CBO9780511610721.005.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=YXogl4uQUO.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ehfRiF0R3a.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=h0ZfDIrj7T.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A MULTI-COMPARISON CORRECTION

Bonferroni and Benjamini–Hochberg corrections were applied jointly across the family of paired tests reported in §6. The family consists of:

- Multi-call diversity (Vendi vs MoA, vs SC): 2 tests.
- Response-level diversity vs single-call baselines (CoT, Thinking) on 4 metrics each: 8 tests.
- Cross-turn drift, Reverie vs A3 BareThinkers and vs A1 NoMemEmo, on 3 windows each (ALL, EARLY, LATE): 6 tests.
- Scale ablation, 6 pairwise contrasts across {3b, 7b, 14b, 32b} on 3 metrics (within-turn Vendi, pairwise SBERT, cross-turn drift): 18 tests.
- Naturalness stable-win binomials (Reverie vs SC, vs MoA): 2 tests.
- Five-rater preference study, GEE test: 1 test.

Total family size $N = 37$. Table 3 lists the headline tests under both corrections.

Table 3: Headline tests under Bonferroni correction at $N = 37$; each row is Reverie vs the named condition. NoMemEmo drift was significant under our original (incorrect) Wilcoxon aggregation; under the corrected per-topic test it is no longer significant under Bonferroni and is reported as such in §6.4.

Test	Raw p	Bonferroni p	BH-FDR p	Status
MoA Vendi	1.77×10^{-6}	6.5×10^{-5}	4.1×10^{-6}	***
SC Vendi	1.44×10^{-14}	5.3×10^{-13}	5.3×10^{-13}	***
BareThinkers drift (ALL)	1.6×10^{-7}	6.0×10^{-6}	4.6×10^{-7}	***
BareThinkers drift (LATE)	1.0×10^{-8}	3.9×10^{-7}	3.2×10^{-8}	***
BareThinkers drift (EARLY)	4.3×10^{-4}	0.016	9.3×10^{-4}	*
NoMemEmo drift (ALL)	0.028	1.00	0.045	n.s. (Bonf)
SC human-like	7.5×10^{-11}	2.8×10^{-9}	4.6×10^{-10}	***
MoA human-like	0.093	1.00	0.132	n.s.
Five-rater preference (GEE)	1.7×10^{-7}	6.4×10^{-6}	4.6×10^{-7}	***

The headline cognitive-substrate findings (within-turn Vendi vs MoA/SC, cross-turn drift vs the no-substrate ablation, naturalness vs SC, five-rater human preference) all survive Bonferroni correction at $p < 10^{-3}$ to $p < 10^{-12}$. The naturalness vs MoA test is not significant under either correction; the NoMemEmo drift test is significant under FDR control ($q = 0.045$) but not under Bonferroni, and is reported as such in §6.4.

B RESPONSE-LEVEL DIVERSITY VS SINGLE-CALL GEMINI BASELINES (PER-METRIC)

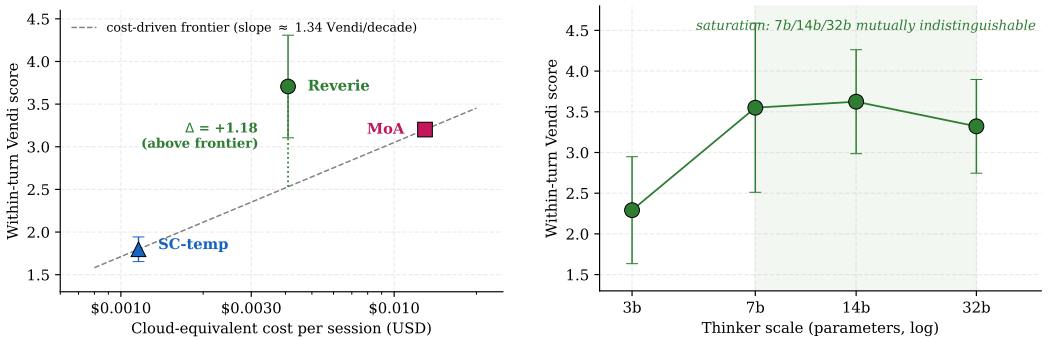
Headline result in §6.2: Reverie wins every length-controlled diversity metric against both single-call Gemini baselines. Table 4 reports the underlying per-metric values, $\Delta\%$ vs Reverie, and uncorrected p-values (paired Wilcoxon, $n = 40$ matched topics). All eight contrasts that are significant in the headline survive Bonferroni correction over the 37-test family (§A); the *Distinct-1 vs Thinking* entry is the one length-confounded exception (Reverie is 44% longer than Thinking, and Distinct-1 is not length-normalised).

C COST AND SCALE

Cost. We charge every condition's LLM output at cloud-equivalent per-token rates (\$0.16–\$5/M; full rate sheet and call counts in Table 5). Reverie issues $\sim$ 175 LLM calls per session vs MoA's 60 (5 per turn), but on cheaper models, and Pareto-dominates MoA: $+15.8\%$ Vendi at $\sim 3\times$ lower per-session cost (Figure 4). SC is cheapest in absolute terms but its diversity collapses (Vendi 1.80

Table 4: Per-metric breakdown for §6.2. Reverie vs Gemini-CoT (single call, $t = 0.7$, thinking budget $= 0$) and Reverie vs Gemini-Thinking (single call, $t = 1.0$, thinking budget $= 1024$). Length-normalised MTLD and the semantic / distributional metrics favour Reverie under both length regimes (Reverie is 18% shorter than CoT, 44% longer than Thinking).

		vs Gemini-CoT			vs Gemini-Thinking		
Metric	Reverie	CoT	Δ%	p	Thinking	Δ%	p
MTLD (length-norm)	150.6	127.7	+17.9	$<10^{-9}$	139.1	+8.3	5×10^{-4}
Distinct-1	0.568	0.504	+12.6	$<10^{-9}$	0.626	−9.4	$<10^{-9}$
SBERT pairwise dist.	0.497	0.467	+6.5	1.3×10^{-3}	0.417	+19.3	$<10^{-9}$
SBERT dist. from prior	0.470	0.447	+5.2	4.9×10^{-3}	0.403	+16.6	$<10^{-9}$
Vendi (session)	5.33	4.96	+7.5	6×10^{-4}	4.44	+20.1	$<10^{-9}$
Tokens (length)	478	583	−18.0	$<10^{-9}$	333	+43.7	$<10^{-9}$



(a) Cost–diversity frontier ($n = 40$, ±1 SD). Reverie sits +1.18 Vendi above the SC-to-MoA cost-driven frontier and Pareto-dominates MoA: +15.8% Vendi at ∼ 3× lower cost.

(b) Scale ablation ($n = 40$ per scale, ±1 SD). Within-turn Vendi saturates at 7b: 7b, 14b and 32b are not separable after correction.

Figure 4: **Cost and scale of the Reverie thinker substrate.**

vs 3.71). Reverie thinkers run on local Qwen GPUs in our setup; the cloud-equivalent figure is what they would cost via a cloud Qwen endpoint. Wall-clock latency (∼ 40 vs ∼ 10 sec/turn) is discussed in §8.

Scale. Varying Reverie's thinker model across {3b, 7b, 14b, 32b} with executive held at qwen2.5:32b (full per-scale means in Table 6): within-turn Vendi saturates at 7b (3b sits 30–36% below the cluster, all three contrasts Bonferroni $p \leq 2.5\times10^{-5}$; 7b, 14b and 32b are not separable after correction, the closest pair being 14b vs 32b at raw $p = 0.02$). The headline result is on cross-turn drift: *none* of the 6 pairwise comparisons reaches significance ($|\Delta| \leq 5\%$, all Bonferroni $p = 1.00$). Drift is scale-invariant — the architectural mechanism, not the per-call thinker capability, is what reduces drift.

Table 5: Cost vs diversity, cloud-equivalent, per 12-turn session. Reverie Pareto-dominates MoA (more diverse, ∼ 3× cheaper); SC is cheapest but its diversity collapses to ∼ 1.8 effective items.

Cond.	Substrate $	Exec $	Total $/sess.	Vendi	$ / Vendi-unit
Reverie	0.00168	0.00241	**0.00409**	**3.71**	**0.00110**
MoA	0.01195	0.00102	0.01297	3.20	0.00405
SC	0.00044	0.00072	0.00116	1.80	0.00064

Table 6: Scale ablation, per-scale means. Within-turn Vendi saturates at 7b (Vendi $\geq$ MoA's 3.20 and SC's 1.80 from §6.1); cross-turn drift is invariant across scales.

Scale	Within-turn Vendi	Pairwise SBERT dist.	Cross-turn drift
3b	2.29 ± 0.66	0.23 ± 0.07	0.30 ± 0.06
7b	3.55 ± 1.04	0.41 ± 0.07	0.30 ± 0.06
14b	3.62 ± 0.64	0.40 ± 0.09	0.31 ± 0.07
32b	3.32 ± 0.58	0.39 ± 0.06	0.30 ± 0.07

D EMBEDDER ROBUSTNESS (MPNET REPLICATION)

The corrected family used only `all-MiniLM-L6-v2` (384-dim, BERT-based). MPNet (`all-mpnet-base-v2`, 768-dim) runs are reported here as a separate-embedder check rather than a new correction family (Table 7).

Table 7: Embedder robustness: $\Delta\%$ effect sizes under MiniLM vs MPNet. "thought" = per-turn proposer set; "response" = exec output. "[n/r]" = result not in the MiniLM file. Effect sizes are within $\pm 10\%$ across embedders, often larger under MPNet (richer 768-dim space).

Comparison	Metric	MiniLM $\Delta\%$	MPNet $\Delta\%$
Reverie vs CoT	Distinct-1 (response)	$+12.59$	$+12.59$
Reverie vs CoT	MTLD (response)	$+17.90$	$+17.90$
Reverie vs CoT	SBERT-from-prior (response)	$+5.20$	$+9.97$
Reverie vs CoT	SBERT pairwise (response)	[n/r]	$+10.43$
Reverie vs CoT	Vendi (response)	$+7.46$	$+10.68$
Reverie vs Thinking	MTLD (response)	$+8.30$	$+8.30$
Reverie vs Thinking	Vendi (response)	$+20.10$	$+21.80$
Reverie vs Thinking	SBERT pairwise (response)	$+19.30$	$+22.50$
Reverie vs BareThinkers	Within-turn Vendi (thought)	$+22.50$	$+20.50$
Reverie vs BareThinkers	Cross-turn drift (thought)	-31.40	—
Reverie vs BareThinkers	Within-turn Vendi (turns 1-4)	$+44.70$	$+41.40$
Reverie vs NoMemEmo	Cross-turn drift (thought)	-14.20	—
Reverie vs NoCritCensor	ALL response-diversity	n.s.	n.s.

All headline cognitive-substrate findings replicate. The NoCritCensor null result also replicates. Lexical findings (MTLD, Distinct-n) are embedding-free.

E ABLATION DECOMPOSITION (FULL PANEL)

For each metric in the main paper we decompose Full Reverie minus each of the three internal ablation conditions. The decomposition reads as a gradient of architectural removal:

- Full Reverie $-$ A1 (NoMemEmo) $\rightarrow$ contribution of K-line memory and emotion-modulated cognition.
- Full Reverie $-$ A2 (NoCritCensor) $\rightarrow$ contribution of internal quality control.
- Full Reverie $-$ A3 (BareThinkers) $\rightarrow$ contribution of the entire Reverie architecture (substrate + agents).

Pattern observed in the ablation suite:

- The Full $\rightarrow$ A1 (NoMemEmo) gap accounts for the bulk of the cross-turn coherence advantage. Under the corrected per-topic Wilcoxon aggregation, the no-substrate (A3 BareThinkers) drift gap is -31.4% pooled / -26% to -35% across windows; removing memory and emotion alone (A1 NoMemEmo) shrinks the effect to -14% pooled (raw $p = 0.028$,

not stable across windows). The substrate itself does most of the work; memory and emotion contribute marginally.

- The Full → A2 (NoCritCensor) gap is null on response diversity under both embedders — critic and censor do not measurably affect the surface form of executive responses in the auto-conversation regime. §7 of the main paper engages this honestly: the critic / censor are designed for the autonomous regime — visible qualitatively in Demo 2 of §F but not quantitatively in the main response-level metrics.
- The Full → A3 (BareThinkers) gap is the full $+22.5\%$ Vendi / -31.4% drift effect — the joint contribution of substrate + all social agents.

F AUTONOMOUS-MODE DEMOS AND N-MATCHED DISCUSSION

Both sessions below are real human interactions with the running Reverie system (not auto-conversation), preserved with the full `events.jsonl` trace.

Demo 1 — 2026-04-20_00-11-03 ("how are you?"). Duration 14.2 min. User messages: 1. Executive responses: 1. Internal nemes: 13 thoughts + 3 critiques (5 agents contributing). Initiative fires: 0. Maps to: "internal cognition before speaking." A trivial greeting elicits substantive internal deliberation; the resulting answer is non-generic.

Demo 2 — 2026-04-20_08-43-10 ("is silence really a virtue?"). Duration 368.1 min (6.1 hr). User messages: 1. Executive responses: 8 (one to user + seven autonomous initiative fires). Internal nemes: 260 thoughts + 60 critiques + 17 memory recalls. Per-agent thought counts: analytical = 83, creative = 64, skeptical = 64, empathetic = 49. Initiative fires: 7 (peak pressure 0.920). K-lines: `episode_tick_20` captured + retrieved 17×. Total ticks: 10,365. Maps to: (a) cognitive duality in the autonomous regime — 8 distinct framings of silence, each on-topic but each opening a new conceptual angle; (b) autonomous initiative — 7 fires from one user input over 6 hours, threshold-0.7 mechanism in operation; (c) K-line memory — `episode_tick_20` retrieved 17× across the 6-hour arc; (d) continuous cognition — 260 thoughts across 4 cognitive styles in one session.

We present autonomous mode as a qualitative demonstration only (the two demos above), not as a quantitative diversity finding.

G MULTI-AGENT DEBATE, INTENTIONALLY NOT INCLUDED

Multi-Agent Debate (MAD; Du et al., 2024) is built around iterative peer revision over K rounds: each of N agents answers in round 1, then in rounds $2..K$ each agent is shown the others' answers and revises. The mechanism drives a population of answers toward *convergence* on a verifiable target. The original benchmarks are GSM8K, MATH, MMLU, biographical factuality — all tasks with a ground-truth answer.

Our open-ended dialogue benchmark has no ground-truth answer to converge toward, and our headline finding is the *opposite* of convergence — within-turn diversity (Claim 1 of the main paper) measures effective distinctness of proposers, which a debate-and- revise round would explicitly reduce.

Compute considerations: MAD with $N = 3$ agents $\times$ $K = 3$ rounds is 9 aggregator-side calls per turn at full Claude/Gemini prices, which would exceed the cloud-equivalent cost of every condition in our study. The implemented baselines (single-call CoT, single-call extended thinking, MoA, SC) span ensemble-style, sampling-style, and single-call alternatives and cover the architectural-recipe hypothesis space relevant to open-ended dialogue diversity.

H FULL HYPERPARAMETERS PER CONDITION

All Gemini calls use model id `gemini-2.5-flash`, pinned. All local Qwen calls use `qwen2.5:32b` (Q4_K_M quantisation) via Ollama unless noted. Table 8 lists the per-condition settings.

Table 8: Full per-condition hyperparameters. "temp" = sampling temperature; "max" = max_output_tokens; "tb" = thinking_budget.

Tag	Name	Configuration
B1	Gemini-CoT	temp $= 0.7$, max $= 1024$, tb $= 0$, CoT system prompt.
B2	Gemini-Thinking	temp $= 1.0$ (Gemini API requires this when extended thinking enabled), max $= 1024$, tb $= 1024$.
B3	Gemini-HighT	temp $= 1.0$, max $= 1024$, tb $= 0$. Tests diversity-via-temperature directly.
M1	MoA	4 proposers at temp $= 0.7$, max $= 150$, isolated; Gemini-2.5-Flash aggregator (temp $= 0.7$, max $= 1024$). Proposer models and the other two MoA configurations in Appendix L.
M2	SC	4 Qwen-32B proposers at temperatures $\{0.5, 0.7, 0.7, 1.0\}$ (mean 0.725, chosen to match Reverie's thinker mean and remove temperature as a confound), generic prompt, max $= 150$, isolated. Aggregator: Gemini-2.5-Flash (same as M1).
F	Reverie	4 thinkers, persona prompts (analytical, creative, skeptical, empathetic), per-call temp $\sim \mathcal{U}[0.5, 0.9]$, max $= 150$. Critic, censor, memory, emotion, initiative, perceiver: see §I for tick rates. Executive: Gemini-2.5-Flash, temp $= 0.7$, max $= 1024$, EXECUTIVE_WAIT_TICKS $= 20$ (≈ 40 s).
A1	NoMemEmo	Reverie with memory + emotion removed; thinker emotion-modifier prompt disabled.
A2	NoCritCensor	Reverie with critic + censor removed.
A3	BareThinkers	Non-Reverie. No nemes / blackboard / tick loop / event bus. Per turn: 4 parallel Qwen-32B calls (one per cognitive style, persona prompts as F, per-call temp $\sim \mathcal{U}[0.5, 0.9]$, max $= 150$) $\rightarrow$ society_context block $\rightarrow$ single Gemini-2.5-Flash exec call (temp $= 0.7$, max $= 1024$, no multi-turn history).
Scale ablation:		
	Scale-3b/7b/14b/32b	Reverie with thinker model varied across {3b, 7b, 14b, 32b}, executive held at qwen2.5:32b. EXECUTIVE_WAIT_TICKS tuned per scale to equalise thoughts/turn density: 3b $\rightarrow$ 2, 7b $\rightarrow$ 3, 14b $\rightarrow$ 15, 32b $\rightarrow$ 20.
Autonomous-mode (§F):		
F-Auto	Reverie-Autonomous	Full Reverie, 1 user seed, no further user input, initiative-driven response cadence.

Sharding: 4 dedicated Ollama daemons each pinned to one consumer GPU (10 topics per shard). Wall-clock: Reverie internal ablation suite $\sim$ 4–5 hr; MoA $\sim$ 2 hr; SC $\sim$ 3 hr; scale ablation $\sim$ 23 hr.

I Architecture details (per-agent specifications)

Blackboard internals. Per-neme fields: type, content (string), salience (float, $[0, 1]$, set at creation), activation (float, decays 2% per tick), time-to-live (TTL or ticks), metadata (dict). Effective attention weight = activation $\times$ salience. Garbage collection at activation < 0.01. Asyncio-locked Python dictionary. Structured non-decaying state: conversation history (last N turns), emotion (valence $\in [-1, 1]$, arousal $\in [0, 1]$, dominant label), initiative pressure (float), active K-line IDs.

Table 9: Per-agent tick rates and roles.

Agent	Tick interval	Role
Perceiver	1	Rule-based; extracts topics / entities / tone from user input; posts PERCEPT nemes (high salience, short TTL).
Analytical thinker	3	Logical / structured persona; per-call temp $\sim \mathcal{U}[0.5, 0.9]$, max $= 150$.
Creative thinker	4	Analogical / lateral persona.
Skeptical thinker	4	Critical / devil's-advocate persona.
Empathetic thinker	5	Human-centred / emotional persona.
Critic	5	Qwen-32B; scores recent THOUGHT nemes on relevance / novelty / coherence (0–1 each), feeds back into blackboard attention weights.
Censor	2	Rule-based; suppresses near-duplicates, low-coherence nemes, stale un-endorsed content.
Emotion	2	Russell valence-arousal (Russell, 2003); decay 5% per tick toward baseline (valence 0, arousal 0.3); injected into thinker prompts.
Memory	4	Rule-based (no LLM call). Retrieval scores K-lines as $0.7 \cdot J + 0.3 \cdot (1 - d_e)$ where J is the Jaccard word-overlap between the K-line's trigger context and the current percept and d_e is the mean absolute valence-arousal difference; top-2 above similarity 0.2 are retrieved. Capture snapshots agent activation counts, top high-salience nemes, and current emotion after a successful exchange.
Initiative	5	Five-signal pressure formula: uncommunicated thoughts above salience threshold; elapsed time since last user message; current arousal; active K-line presence; high-salience uncommunicated creative thoughts. Threshold 0.7, 60-s cooldown.
Executive	event-driven	Gemini-2.5-Flash. Subscribes to USER_MESSAGE and INITIATIVE_THRESHOLD. `asyncio.Lock` prevents concurrent execution. Waits EXECUTIVE_WAIT_TICKS $= 20$ (≈ 40 s) before reading top-10 THOUGHT + top-5 CRITIQUE + top-3 MEMORY_RECALL nemes.

PER-AGENT SPECIFICATIONS

The remainder of this section documents each agent in enough detail to reproduce. Table 9 above is the quick-reference summary; the paragraphs below give the full specification (inputs, outputs, formulas, prompt structure, hyperparameters).

Perceiver. Rule-based text analyser; tick interval 1. Reads the highest- salience `USER_INPUT` neme (salience > 0.8), processes it once (deduplicated by neme id), and emits one `PERCEPT` neme (salience $= \max(0.5, u)$, TTL $= 80$ ticks). The percept content is the first 150 characters of the user message, prefixed with "[Question]" when a question mark is present; metadata includes *topics, entities, questions, emotional_tone, urgency*. Tone is determined by counting overlap with two fixed word sets (~ 20 words each; positive $= \{$great, good, love, thanks, $\dots\}$, negative $= \{$bad, terrible, hate, problem, $\dots\}$) and is set to "positive" if $|\text{pos}| > |\text{neg}|$, "negative" if $|\text{neg}| > |\text{pos}|$, else "neutral". Urgency $u = \min(0.3 + 0.2 \cdot |U|, 1.0)$ where U is the urgency word set ($\{$urgent, asap, immediately, critical, deadline, $\dots\}$). Topics are length-> 3 words outside the stopword set (top 5); entities are capitalised words from the original casing (top 5).

Thinkers (analytical, creative, skeptical, empathetic). LLM-based; tick intervals $\{3, 4, 4, 5\}$. Each thinker has a distinct system prompt fixing its cognitive style: analytical breaks problems into components and looks for causal chains; creative thinks by analogy and metaphor and asks "what if?"; skeptical challenges assumptions and plays devil's advocate; empathetic considers the human

and emotional dimension. Per-call sampling temperature is drawn fresh from $\mathcal{U}[0.5, 0.9]$; max output $= 150$ tokens.

Active mode. When the freshest `PERCEPT` has activation > 0.868 (~ 14 s after creation), the thinker reads up to 3 percepts, 5 peer thoughts, and 2 memory recalls, formats them with the emotional modifier (next paragraph), and queries the LLM. Output is a `THOUGHT` neme (salience computed from text properties, TTL $= 40$).

Wandering mode. When no fresh percepts are present, the thinker generates an autonomous thought. Firing is staggered to avoid all four wandering at once: only every 12 ticks per style with offsets $\{0, 3, 6, 9\}$. If a user message has been seen, the wandering is *anchored* (uses an anchored prompt with the active question plus full blackboard context); otherwise it is *free* (no anchoring topic; the prompt shows raw valence/ arousal numbers directly).

Emotional modifier (used in active and anchored-wandering prompts): a short prose instruction selected by valence-arousal quadrant. For example, $v > 0.3$, $a > 0.5 \Rightarrow$ "You are feeling engaged and optimistic. Think expansively."; $v < -0.3$, $a > 0.5 \Rightarrow$ "You sense something is wrong. Think carefully and critically.". Free-wandering prompts inject the raw numbers (`valence=<v>, arousal=<a>`) instead, since there is no anchoring topic for a tone instruction to attach to.

Critic. LLM-based (Qwen-32B); tick interval 5. Reads up to 10 recent `THOUGHT` nemes that have not yet been critiqued, picks one, and prompts the LLM to score it on three dimensions $r, n, c \in [0, 1]$ (relevance, novelty, coherence) plus a one-sentence critique, with JSON-only output. The target thought's salience is updated as

$$s' = 0.5 \cdot s + 0.5 \cdot \bar{c}, \qquad \bar{c} = (r + n + c)/3,$$

so effective attention on subsequent reads becomes activation $\times s'$. The thought is tagged "critiqued" (so it will not be re-rated) and the scores are attached to its metadata for downstream use by the initiative agent. Emits a `CRITIQUE` neme (salience 0.4, TTL 30) carrying the critique text.

Censor. Rule-based; tick interval 2. Reads up to 10 recent `THOUGHT` nemes (including already-suppressed ones, so a reason can be logged). Suppresses if any of: content length < 5 chars; content hash matches one of the last 50 thoughts (literal duplicate); content contains "as an AI" or "as a language model" (LLM-degeneration markers); salience < 0.1. On suppression, marks the thought id in `blackboard.suppressed_ids` (so other agents skip it) and emits a `SUPPRESSION` neme (salience 0.6, TTL 20) recording the reason.

Memory. Rule-based (no LLM call); tick interval 4. Operates in two modes per tick.

Retrieval: reads the top 2 `PERCEPT` nemes; if any exist, scores every K-line in the store as

$$\mathrm{sim}(k, q) = 0.7 \cdot J(k, q) + 0.3 \cdot (1 - d_e(k, q)),$$

where J is the Jaccard word-overlap between the K-line's stored trigger context and the joined percept text, and $d_e = (|\Delta v| + |\Delta a|)/2$ is the mean absolute valence-arousal difference. The top 2 K-lines with similarity ≥ 0.2 are retrieved; their ids become `blackboard.active_k_lines` (read by the initiative agent) and a `MEMORY_RECALL` neme is emitted (salience 0.7, TTL 30) carrying the K-lines' neme summaries.

Capture: after every assistant response (heuristic: the last two conversation-history entries are user $\rightarrow$ assistant and the 10-tick capture cooldown has elapsed), snapshot the current cognitive state as a new K-line. The K-line stores: per-agent activation rate (activations / tick), the top 5 high-salience (≥ 0.4) neme contents, the recent percept text as trigger context, and the current emotion vector. The episodic store is bounded to the most recent 500 episodes.

Emotion. Computational (no LLM); tick interval 2. Maintains valence $v \in [-1, 1]$ and arousal $a \in [0, 1]$ on the blackboard as a Russell-circumplex (Russell, 2003) state vector. Each tick:

(1) Compute deltas. For each recent percept (≤ 3): $d_v \mathrel{+}= +0.15$ if its tone is positive, -0.15 if negative; $d_a \mathrel{+}= 0.2 \cdot s_p$. For each thought with critic scores (≤ 5): $d_v \mathrel{+}= 0.1 \cdot (\bar{c} - 0.5)$. For each suppression ($\leq 3$): $d_a \mathrel{+}= 0.05$. If ticks-since-last-user-input > 20 (~ 40 s), $d_a \mathrel{-}= 0.1$ and $d_v \mathrel{*}= 0.9$.

(2) Apply momentum (leaky integrator, $\rho = 0.7$): $\mathrm{vm} \leftarrow \rho \cdot \mathrm{vm} + d_v$; $\mathrm{am} \leftarrow \rho \cdot \mathrm{am} + d_a$.

(3) Update state and decay ($\eta = 0.05$): $v_{\text{new}} = (v + \text{vm}) \cdot (1 - \eta)$; $a_{\text{new}} = (a + \text{am}) \cdot (1 - \eta) + \eta \cdot 0.3$. Clip $v \in [-1, 1]$, $a \in [0, 1]$.

(4) Re-classify the discrete label by quadrant: *excited, content, anxious, melancholy, alert, calm, neutral* (boundaries $|v| > 0.3$ and a thresholds at 0.5, 0.6, 0.3). Emits an `EMOTION_SIGNAL` neme (salience 0.3, TTL 10).

Initiative. Threshold logic (no LLM); tick interval 5. Each tick computes a scalar pressure

$$\begin{aligned} P = \; & \min(0.08 \cdot N_u,\ 0.25) + 0.12 \cdot N_h \\ & + \min(0.005 \cdot \max(0,\ T - 30),\ 0.2) + 0.15 \cdot a \\ & + 0.10 \cdot \mathbb{1}[K] + 0.10 \cdot \mathbb{1}[C], \end{aligned}$$

where N_u is the number of uncommunicated thoughts with salience > 0.3 (capped contribution 0.25); N_h is the count of those with mean critic score > 0.7; T is ticks since the last user message; a is current arousal; $\mathbb{1}[K]$ is 1 if any K-line is active; $\mathbb{1}[C]$ is 1 if any uncommunicated thought from the creative thinker has salience > 0.5. P is clipped to $[0, 1]$ and written to `blackboard.initiative_pressure`. When $P \geq 0.7$ and the 60-s cooldown has elapsed, the agent publishes an `INITIATIVE_THRESHOLD` event (which the executive subscribes to) and records the firing tick to start the cooldown.

Executive. LLM-based (cloud or local); event-driven (tick interval 0, so the tick loop skips it). Subscribes to `USER_MESSAGE` and `INITIATIVE_THRESHOLD` events; an `asyncio.Lock` prevents concurrent executions. On trigger, sleeps for `EXECUTIVE_WAIT_TICKS` $= 20$ ticks (~ 40 s) so thoughts can accumulate, then reads the current state: top 10 `THOUGHT` nemes (salience ≥ 0.3), top 5 `CRITIQUE`, top 3 `MEMORY_RECALL`, plus the emotion vector. Formats the state as a prose "society context" string (no section headers; an earlier header-style format caused `gemini-flash` to echo the prompt verbatim into autonomous responses). Calls the LLM with this system prompt as a single Q→A call — no multi-turn conversation history is passed; cross- turn state is carried entirely by the substrate (continuous thinking, K-line memory). On autonomous fires, the call includes an internal trigger message indicating the initiative threshold was crossed. Sampling temperature is 0.7 for user-triggered responses, 0.9 for autonomous fires (to break repetition attractors); max output 1024 tokens. After the call, marks every read thought as "communicated" (so it does not re-trigger the initiative agent), appends the response to conversation history, and emits an `EXECUTIVE_OUTPUT` neme (salience 0.9, TTL 100). A secondary baseline call (same prompt with the society-context block replaced by "(No internal cognitive activity yet.)") is also issued to compute a *society influence* score per response, stored in metadata for downstream analysis and not visible to the user.

Active-vs-wandering threshold for thinkers (relevant here): percept activation > 0.868 (≈ 14 s from creation). Wandering thoughts have salience 0.3 and shorter TTL.

J HUMAN EVALUATION: PROTOCOL, RUBRICS, AND QUESTIONS

This appendix documents the human evaluation referenced in the main paper.

HUMAN PREFERENCE, 60 QUESTIONS × 5 RATERS (SECTION 6.3)

Stimuli. 60 open-ended questions, listed in Table 10. Each was run as a 12-turn conversation with the simulated user of §4 under two systems: Reverie (default configuration, Gemini-2.5-Flash executive, 40 s wait) and MoA in the Experiment 2 configuration of Appendix L.

Summaries. Raters judged summaries, not transcripts. Each 12-turn conversation was condensed to at most 200 words by `gemini-2.5-flash`.

Raters and blinding. Five raters unaffiliated with the project, blinded to system identity and to the hypothesis. Each received a package of 60 pairs with the order of questions and the A/B assignment shuffled from a per-rater seed, and answered every pair. The instruction was to pick the better overall response, the one that is more substantive, insightful and useful.

Statistics. The per-rater counts are binomial; the pooled estimate uses a Bayesian mixed-effects logistic model with a rater random intercept (variational fit) and, as a frequentist check, a binomial

GEE with an exchangeable working correlation and cluster-robust standard errors by rater. Item-level agreement is Fleiss κ over the five raters (0.043); a majority of raters chose Reverie on 47 of 60 items.

THE 60 OPEN-ENDED RESEARCH QUESTIONS

All 60 questions below were used in the preference study (each rated by all five raters, with order and A/B assignment shuffled per rater), given verbatim as the opening user message of each 12-turn conversation.

Table 10: The 60 open-ended research questions used in the preference study.

#	Question
1	Do black holes really exist as Einstein described them? And if they do, could we ourselves be inside one — could the entire observable universe be the interior of a black hole? Think this through carefully.
2	Is consciousness substrate-dependent, or could a sufficiently complex non-biological system be conscious? How would we even tell? Reason carefully.
3	Is room-temperature superconductivity at ambient pressure physically possible, or is it fundamentally precluded by the laws of physics? Think through the constraints.
4	Why is the brain 'immune-privileged' — protected from much of the body's immune response? What does that say about the relationship between identity and immunity? Think through the implications.
5	Is P vs NP a question about computation, or is it a question about the structure of mathematical truth itself? What would each answer mean? Think carefully.
6	Is the wave function in quantum mechanics a real physical thing, or just a mathematical bookkeeping device for our knowledge of the world? Think this through carefully.
7	Is free will real, or is the felt experience of choosing an illusion produced by the brain after the fact? Would knowing the answer change how we live? Think this through.
8	Is matter ultimately continuous, or is the universe granular at some smallest possible scale? What hangs on the answer? Think this through carefully.
9	Why does the body sometimes attack itself? Is autoimmunity a malfunction, or is it a deep feature of any sufficiently sophisticated immune system? Think carefully.
10	Are there computations that are genuinely impossible — not merely hard — and what would ground that limit? Think carefully through what 'impossible' means here.
11	Does time actually flow, or is the sense of a moving 'now' an artifact of how brains process memory? If flow is an illusion, why is it so universal? Reason carefully.
12	Are viruses alive? And does the fact that the question is so hard to answer reveal something wrong with our concept of life itself? Reason carefully.
13	Why do we sleep and dream at all, when both leave an animal vulnerable for hours? Is dreaming a function or a side effect? Reason carefully through both possibilities.
14	Is memory a recording of the past, or a reconstruction rebuilt each time we recall? If reconstruction, in what sense is any memory true? Reason carefully.
15	Is the immune system best understood as a defense, or as an organ of self-definition that decides what counts as 'me'? What follows from each view? Reason carefully.
16	Is the universe finite or infinite? If it is finite, what could possibly be 'outside' of it? Reason carefully through the conceptual difficulties.
17	Is Earth's climate fundamentally chaotic? If so, are long-term predictions impossible in principle, or only in practice? Think carefully about what 'predictability' means here.
18	Is dark matter actually a particle, or is it a sign that our theory of gravity itself is wrong at large scales? How could we distinguish? Reason carefully.
19	How much of a person is determined at conception versus by environment, and is the question even well-posed? Think carefully about what determinism means in genetics.
20	Can probability be objective — a property of the world — or is it always a statement about belief? What hangs on the answer? Reason carefully.
21	If the constants of nature appear fine-tuned for life, is this evidence of design, of a vast multiverse, or of a selection effect we don't yet understand? Reason carefully.
22	If we successfully geoengineer Earth's climate to counteract warming, could we ever safely stop? What does this say about the kind of commitment such interventions represent? Reason carefully.
23	Could there be life forms based on completely different physics elsewhere in the universe? If so, would we recognize them as alive? Reason carefully.
24	If we could read a person's entire genome perfectly, could we predict who they would become? What does that question even mean? Reason carefully.

continued on next page

Table 10 – continued from previous page

#	Question
25	Can two perfectly rational people see exactly the same data and reach different conclusions? If yes, what does that say about the nature of evidence? Reason carefully.
26	Is mathematics unreasonably effective at describing the physical world, or does it only seem that way because we notice the successes and forget the failures? Think this through carefully.
27	Is the periodic table a discovery about nature, or a human classification imposed on a continuum of matter? What hangs on the difference? Think this through.
28	Could an intelligence far greater than ours discover truths that are not merely hard for us, but permanently beyond human comprehension? How would we recognize our own limit? Think this through.
29	Could life exist that is not based on chemistry at all — in plasma, in fields, in patterns of pure information? Would we have any way to recognize it? Think this through.
30	Could two different physical theories make identical predictions forever, so that no experiment could ever choose between them? What would that mean for scientific truth? Think this through.
31	Is there a fundamental definition of life? One precise enough that we could tell whether a sufficiently complex AI system would count as alive? Think through the implications.
32	Can machines truly understand, or do they only simulate understanding? What would the empirical difference look like — could we ever distinguish them? Think this through.
33	Why does Earth have plate tectonics when other rocky planets in the solar system don't? Is plate tectonics essential for life as we know it? Think carefully through the implications.
34	Is the deep ocean essentially a separate biosphere from the surface, with its own evolutionary trajectory? What would that mean for our understanding of life on Earth? Think carefully.
35	Could intelligent life have evolved on Earth before humans and left no trace? What kind of evidence would even survive across geological time? Think carefully through this.
36	Is biological aging a programmed process, or a slow accumulation of accidental damage? If we could fully reverse it, should we? Think through both questions.
37	Can a language model trained only on text develop genuine concepts about the physical world it has never directly experienced? How would we test this? Think carefully.
38	What would Earth look like if liquid water had never existed on its surface? Would anything like life have emerged at all? Think this through carefully.
39	Could there be entire ecosystems on Earth that depend on no sunlight at all, beyond the deep-sea vents we already know? What would they reveal about the boundaries of life? Think carefully.
40	Mass extinctions punctuate the fossil record. Are they truly catastrophic events, or are they visible markers of slower, deeper restructurings of life? Think carefully.
41	Could a civilization arise that never developed the concept of number? What would its science and technology even look like? Reason carefully through the possibilities.
42	Could there be a largest possible organism, set by physics rather than biology? What limits would actually bind it? Reason carefully through the constraints.
43	Is color a property of objects, of light, or only of minds? If minds, then in what sense does the external world have any appearance at all? Reason carefully.
44	Is extinction always a loss, or can it be a precondition for new forms that could not otherwise arise? What does the answer imply about how we value life? Reason carefully.
45	Why is there so much more matter than antimatter in the universe, and could a mirror cosmos of antimatter exist somewhere we cannot see? Reason carefully through the physics.
46	Why does life on Earth use only L-amino acids? Could mirror-life chemistry — based on D-amino acids — exist somewhere in the universe? What would that imply?
47	Are mathematical truths discovered or invented? What would each answer mean for the relationship between mathematics and physical reality? Reason carefully.
48	Are ecosystems self-organizing systems with emergent stability, or are they just statistical aggregates we project structure onto? What would each view imply for conservation? Reason carefully.
49	Is life on Mars fundamentally impossible, or just hard to detect? What would settle the question? Reason carefully through the possibilities.
50	Is there a fundamental limit to how much energy we can extract from the vacuum? What would it mean if the answer is no? Reason carefully through the physics.
51	What was the first self-replicating molecule? Could it have been something other than RNA, and what would that mean for the origin of life? Reason carefully.
52	Are there mathematical truths that humans will never prove? If so, how would we ever know which ones are unreachable rather than merely difficult? Reason carefully.
53	Are extinct species lost forever, or is there a meaningful sense in which they could return — and what does the answer reveal about identity in biology? Reason carefully.
54	Is the long-term stability of the solar system fragile or robust? Could a small perturbation today eventually unseat a planet? Reason carefully through what we know.

continued on next page

Table 10 – continued from previous page

#	Question
55	Is fusion fundamentally harder than fission, or are we just engineered against it? What would it mean if commercial fusion were physically impossible? Reason carefully.
56	Is pain fundamentally private, or could two beings ever share the same sensation? What would it take to know? Think carefully about what sharing an experience would mean.
57	Is randomness a real feature of the world, or just a name for causes we cannot see? How would we ever tell the two apart? Think carefully.
58	Are the laws of physics the same everywhere and everywhen, or could they drift across cosmic time? What would it take to detect a change? Think carefully through the difficulty.
59	Do numbers exist independently of minds, waiting to be found, or would they vanish if every thinking being disappeared? Think carefully about what 'exist' means here.
60	Is understanding a language the same as knowing how to use it, or is there something more that mere fluency can never capture? How would we test the difference? Think carefully.

K VERIFIABLE TASKS AND SENSITIVITY SWEEP: SETUP DETAILS

Table 11: **Sensitivity sweep** (21 configurations $\times$ 10 topics, 12-turn sessions). Within-turn Vendi at the natural per-turn thought count (mean over sessions) and cross-turn drift (lower = more anchored); p is Mann–Whitney vs. the default on the 10 session means (Vendi). Default = decay 0.02, wait 20 ticks, threshold 0.7, cooldown 30 ticks. Grid-cell drift values were computed in-session and transcribed. Bold: significant at 0.05.

Parameter	Setting	Vendi	p	Drift
Default	(see caption)	3.24	—	0.32
Decay rate	0.01	3.35	0.52	0.33
	0.05	2.64	**0.002**	0.35
	0.10	2.51	**0.0006**	0.37
Executive wait	10 ticks (20 s)	2.41	**0.0002**	0.33
	30 ticks (60 s)	3.65	0.076	0.28
Initiative threshold	0.5	3.25	0.97	0.29
	0.9	3.37	0.38	0.28
Cooldown	15 ticks	3.25	0.91	0.32
	60 ticks	3.24	0.97	0.29
Roles	drop creative	2.83	**0.021**	0.28
	drop skeptical	2.78	**0.011**	0.30
	drop analytical	2.85	n/t	—
	drop empathetic	2.98	n/t	—
	swap roles	3.37	0.73	0.32
Decay$\times$wait grid	0.10, 10 ticks (fast, short)	1.91	**0.0002**	0.39
	0.01, 30 ticks (slow, long)	3.66	**0.038**	0.29

This appendix documents the setup of Experiments 3 and 4 (§6.6–§6.7); the five-rater preference study (§6.3) is documented in Appendix J.

VERIFIABLE TASKS (SECTION 6.6)

Items. 100 items per dataset, drawn once and frozen; the frozen files are released. Datasets: TriviaQA, NQ-Open, PopQA (factual QA); GSM8K, MATH-500, SVAMP (math); HumanEval, MBPP, MBPP+ (code).

Systems. Both systems receive the identical question and output-format instruction. Reverie runs its default configuration with Qwen2.5:32b thinkers and a Gemini-2.5-Flash executive, 40 s wait. MoA runs in the Experiment 3 configuration of Appendix L. Local calls are capped at 150 output tokens and the final answer at 400 for both systems.

Grading. Exact/normalised-match for factual QA, numeric match for GSM8K and SVAMP, `\boxed{}` extraction for MATH-500, and unit tests for the code sets. Planning (PlanBench Blocksworld, Natural-Plan trip and calendar) was run but dropped because prose answer extraction scored near zero for both systems while the gold self-check scored 100%, so the number would have measured formatting.

Statistics. Per dataset, a paired bootstrap over the 100 items ($10{,}000$ resamples) gives a two-sided p and a 95% CI on the accuracy difference. Call and token counts are means over items of the per-item logs (local 32b calls; generated tokens for the local model and for the shared Gemini-Flash executive/aggregator; prompt tokens excluded).

Sensitivity sweep and role ablations (Section 6.7)

A smaller version of the Experiment 1 protocol: 21 configurations, each run on the same 10 seed topics (rather than 40) for 12 turns with the simulated user. The default is (decay 0.02 per tick, executive wait 20 ticks, initiative threshold 0.7, cooldown 30 ticks; one tick $= 2$ s); decay $\in \{0.01, 0.05, 0.10\}$; wait $\in \{10, 30\}$ ticks; threshold $\in \{0.5, 0.9\}$; cooldown $\in \{15, 60\}$ ticks; a 3×2 decay$\times$wait grid ($\{0.01, 0.05, 0.10\} \times \{10, 30\}$); and five role configurations (drop each of the four thinkers; swap roles, in which the skeptical thinker is replaced by a second analytical thinker). Within-turn Vendi follows the L_1 protocol of §5 (per-turn set of thoughts posted between the user message and the executive response, MiniLM, mean over turns) at matched $N = 4$. Cross-turn drift is the turn-1 to turn-t centroid distance averaged over turns. Significance is Mann–Whitney on the 10 session means against the default. Because only 10 topics were used, the default row (3.24) is not directly comparable to the 40-topic value of Figure 3a (3.71). Drift for the two grid extremes was computed in-session and transcribed; drift for the drop-analytical and drop-empathetic configurations was not computed.

Results. Table 11 reports the sweep, scored with the same within-turn Vendi. The default row (3.24) comes from this small sample of 10 topics and is not directly comparable to the 3.71 of Figure 3a. Diversity is the sensitive axis, and only two parameters move it: faster decay lowers Vendi (0.05: 2.64, $p = 0.002$; 0.10: 2.51, $p = 0.0006$) and a shorter executive wait lowers it (10 ticks: 2.41, $p = 0.0002$); in the decay$\times$wait grid, fast decay with a short wait collapses Vendi to 1.91 ($p = 0.0002$) and slow decay with a long wait raises it to 3.66 ($p = 0.04$). These are exactly the two parameters that control how many thoughts accumulate before the executive reads. Initiative threshold and cooldown move neither metric (all $p \geq 0.38$); the threshold changes only how often the system speaks unprompted. Cross-turn drift stays within 0.28–0.37 across every configuration and no drift contrast reaches significance (all $p > 0.24$).

Roles matter. Dropping the creative or the skeptical thinker lowers within-turn Vendi by 13–14% ($p = 0.021$ and 0.011); dropping the analytical or empathetic thinker gives similar means (2.85, 2.98). Swapping roles (skeptical replaced by a second analytical thinker) has no effect (3.37, $p = 0.73$). We do not claim the four roles are cognitively privileged; the evidence supports a small set of complementary styles, each of which contributes.

L Mixture-of-Agents configurations

Mixture-of-Agents (MoA; Wang et al., 2025) has N proposers answer a prompt independently and one aggregator synthesise their answers. The paper uses MoA in three experiments, with the same aggregator throughout and different proposers in each. What is common to all three: the aggregator is `gemini-2.5-flash` with the canonical MoA aggregate-and-synthesise prompt (the same model as Reverie's executive, so the architectural contrast is on the proposer side); every proposer call is capped at 150 output tokens (Reverie's thinker cap); and proposers are stateless per turn, seeing only the current user message, as Reverie's executive does (§3). Table 12 lists the differences.

Experiment 1 used cloud proposers from two model families, Claude Haiku 4.5 and 3.5 (Anthropic, 2025a; 2024) and Gemini 2.0 and 2.5 Flash-Lite (Google DeepMind, 2025). Two of them were subsequently retired by their providers (Claude 3.5 Haiku in February 2026, Gemini 2.0 Flash-Lite in June 2026), so the cloud configuration could not be reused; Experiments 2 and 3 therefore used Reverie's own local thinker model as the proposer, with a 40 s per-turn budget so that the human and

Table 12: The three MoA configurations. All share a `gemini-2.5-flash` aggregator and a 150-token proposer cap.

	Experiment 1 (§6.1, §6.5)	Experiment 2 (§6.3)	Experiment 3 (§6.6)
Proposer models	Claude Haiku 4.5, Claude Haiku 3.5, Gemini 2.0 Flash-Lite, Gemini 2.5 Flash-Lite (one call each)	$4\times$ Qwen2.5:32b (Reverie's thinker model)	$4\times$ Qwen2.5:32b
Proposer prompt	Reverie's four cognitive-style prompts (analytical, creative, skeptical, empathetic)	Reverie's four cognitive-style prompts (analytical, creative, skeptical, empathetic)	generic, plus the task's output-format instruction
Temperature	0.7	0.7	$\{0.3, 0.5, 0.7, 0.9\}$
Aggregator cap	1024	1024	400 (Reverie's executive cap in this experiment)
Setting	12-turn conversations on 40 topics	12-turn conversations on 60 questions	single-shot items from 9 benchmarks

verifiable-task comparisons are matched to Reverie in model, token caps and time. The Experiment configurations are in Table 5.