MIRROR: CONVERGING COGNITIVE PRINCIPLES AS COMPUTATIONAL MECHANISMS FOR AI REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple cognitive theories—Global Workspace Theory, reconstructive episodic memory, inner speech, and complementary learning systems—converge on a shared set of architectural principles: parallel specialized processing, integrative synthesis into a bounded unified representation, and reconstructive rather than accumulative maintenance. We test whether these converging principles provide computational advantages when implemented in AI systems. MIRROR operationalizes each principle as a concrete mechanism: an Inner Monologue Manager generates parallel cognitive threads (Goals, Reasoning, Memory), a Cognitive Controller synthesizes these into a bounded first-person narrative that is fully reconstructed each turn, and a temporal separation between fast response generation and slow deliberative consolidation mirrors complementary learning dynamics. Evaluated on multi-turn dialogue requiring constraint maintenance under attentional interference, MIRROR yields 21% relative improvement across seven architecturally diverse language models. Ablation studies test the theoretical predictions directly: reconstructive synthesis improves all seven models (+5–20%); the integrated system outperforms either component alone for six of seven models, confirming that parallel exploration and integrative synthesis are complementary; and gains concentrate where theories predict—under high attentional load where global availability of integrated information is most needed. These results demonstrate that converging principles from human cognition provide architecture-general computational advantages, and generate testable behavioral predictions about working memory, inner speech, and memory consolidation.

1 INTRODUCTION

Human conversation relies on parallel internal processing—recalling memories, tracking goals, monitoring social dynamics—operating beneath conscious awareness (Pickering & Garrod, 2013; Egorova et al., 2013). What makes this processing effective is not any single mechanism but the *convergence* of multiple cognitive principles: parallel specialized processors synthesized into a unified workspace (Baars, 1988; Dehaene & Changeux, 2011), reconstructive memory that regenerates understanding rather than accumulating traces (Bartlett, 1932; Schacter, 2012), inner speech providing self-regulatory coherence (Morin, 2011; Vygotsky, 1962), and complementary fast and slow learning systems (McClelland et al., 1995; Kumaran et al., 2016).

These principles, developed in separate research traditions, converge on a common architectural signature: *parallel exploration feeding into bounded integrative synthesis, maintained reconstructively through self-referential narrative, with temporal separation between fast response and slow consolidation.* This convergence suggests these may not be independent adaptations but reflections of a deeper computational logic. Yet this convergence remains untested as a unified computational mechanism.

Current AI approaches to multi-turn dialogue implement at most one of these principles in isolation. Chain-of-thought reasoning (Wei et al., 2022) generates deliberation but discards it. Reflexion (Shinn et al., 2023) and MemGPT (Packer et al., 2023) maintain memory but accumulate traces unboundedly. Extended reasoning modes produce rich deliberation within turns but maintain no

Table 1: Converging cognitive principles mapped to MIRROR components and testable predictions

Cognitive Theory	Principle	MIRROR Component	Prediction
Global Workspace Theory	Parallel processors → unified broadcast	Threads → Controller synthesis	Non-additive gains from integration
Reconstructive Memory	Regeneration, not accumulation	O(1) narrative reconstruction	Bounded reconstruction outperforms accumulation
Inner Speech	First-person self-regulation	Consistent self-reference across components	Improved cross-turn coherence
Complementary Learning	Fast response + slow consolidation	Talker (fast) + Thinker (slow)	Benefits from temporal separation

state across them. None implement the full convergent architecture that cognitive science suggests is fundamental to human cognition.

We introduce MIRROR, a cognitive architecture that operationalizes these converging principles as testable computational mechanisms, synthesizing principles that multiple theories independently identify as critical. From **Global Workspace Theory** (Baars, 1988; Dehaene & Changeux, 2011), MIRROR implements parallel cognitive threads whose outputs are integrated into a single globally available representation. From **reconstructive episodic memory** (Bartlett, 1932; Baddeley, 2000; Schacter, 2012), the system fully regenerates a bounded narrative each turn—O(1) reconstruction rather than O(n) accumulation. From **inner speech theory** (Morin, 2011; Vygotsky, 1962; Ben Alderson-Day, 2015), consistent first-person self-reference across components creates unified identity through narrative. From **complementary learning systems** (McClelland et al., 1995; Kumaran et al., 2016), immediate response generation is temporally separated from asynchronous deliberative consolidation.

This convergent framing enables specific predictions: (a) benefits should generalize across architecturally diverse models, reflecting organizational rather than substrate-specific advantages; (b) parallel processing and integrative synthesis should contribute non-additively, confirming they address different computational needs; and (c) gains should concentrate under high attentional load, where global availability of integrated information matters most. Our results confirm all three predictions across seven models, establishing that converging cognitive principles provide genuine computational advantages and generating testable predictions for human cognition research.

2 From Cognitive Theory to Computational Mechanism

Rather than treating cognitive science as metaphorical inspiration, we map specific theoretical claims to concrete computational mechanisms and derive testable predictions. Table 1 summarizes these mappings.

2.1 Global Workspace Theory: Parallel-to-Unified Processing

Global Workspace Theory (GWT) (Baars, 1988; Dehaene & Changeux, 2011) proposes that cognition involves parallel specialized processors operating unconsciously, with selected outputs broadcast to a unified "global workspace" that makes information widely available for reasoning and action. The workspace's power lies not in the parallel processing itself but in the *integration*: binding diverse perspectives into a single coherent representation that can guide downstream behavior.

Operationalization. MIRROR's Inner Monologue Manager generates three parallel cognitive threads within a single inference call, each tracking a distinct dimension: **Goals** (user objectives, constraint conflicts), **Reasoning** (logical patterns, causal relationships), and **Memory** (user-specific information, preferences, critical constraints). The Cognitive Controller then synthesizes these parallel outputs into a single bounded representation—the "global broadcast" that becomes available to the Talker for response generation.

Prediction. If the parallel-to-unified pipeline provides genuine advantages (as GWT claims), the integrated system should outperform either parallel threads alone or synthesis alone—a non-additive, synergistic gain reflecting complementary computational functions.

2.2 Reconstructive Episodic Memory: Regeneration Over Accumulation

A foundational finding in memory research is that human recall is not reproductive playback but active reconstruction (Bartlett, 1932). Schacter's constructive episodic simulation hypothesis (Schacter, 2012) argues this reconstructive nature is *adaptive*: by regenerating understanding each time, the system can flexibly integrate new information with prior knowledge rather than rigidly replaying stored traces. Baddeley's episodic buffer (Baddeley, 2000) serves this integrative function within working memory, binding information from multiple subsystems into coherent episodes. Memory consolidation research (Dudai, 2004; Squire & Dede, 2015) further demonstrates that post-encoding processes—not encoding itself—determine what is retained.

Operationalization. MIRROR's Cognitive Controller fully regenerates a bounded first-person narrative ($\leq$3k tokens) each turn, discarding the previous version. Critically, the Controller has no access to raw conversation history—only thread outputs and the prior narrative—forcing genuine information compression rather than passive copying. This implements O(1) bounded reconstruction: regardless of conversation length, the internal representation remains fixed-size.

Prediction. If reconstructive synthesis is the primary advantage (as memory theory suggests), the Cognitive Controller should provide consistent gains across diverse models, and should outperform trace-accumulation approaches that grow unboundedly.

2.3 Inner Speech: Self-Regulation Through Narrative

Research on inner speech (Morin, 2011; Ben Alderson-Day, 2015; Vygotsky, 1962) demonstrates that humans maintain continuous self-directed narrative for planning, self-regulation, and metacognitive monitoring. Vygotsky's developmental account traces inner speech to internalized social dialogue that serves a fundamentally regulatory function. Morin (Morin, 2005) argues that inner speech enables coherent behavior across time by maintaining a persistent self-model.

Operationalization. MIRROR maintains consistent first-person self-reference across all components through a role-based framework: the Talker operates as "the voice," the Inner Monologue Manager as "the subconscious," and the Cognitive Controller as "the core awareness" of a unified system. The reconstructed narrative is maintained in first-person voice ("I understand that the user has..."), creating coherence through narrative identity rather than parameter sharing.

Prediction. If first-person framing serves a genuinely regulatory function, systems maintaining consistent self-referential narrative should show improved coherence compared to third-person or unframed alternatives.

2.4 Complementary Learning Systems: Fast and Slow Processing

Complementary learning systems theory (McClelland et al., 1995; Kumaran et al., 2016) proposes that effective cognition requires both fast, adaptive systems (hippocampus) and slow, consolidative systems (neocortex) that serve complementary functions. Rapid learning captures episodic details; slow consolidation extracts generalizable patterns. Offline processing during rest periods enables consolidation without interfering with real-time behavior (Dudai, 2004).

Operationalization. MIRROR separates the Talker (fast, immediate response generation) from the Thinker (slow, asynchronous deliberative processing). At turn $t = 0$, the Talker responds immediately without internal narrative; the Thinker begins processing after response delivery. For subsequent turns, the Talker uses the previous turn's narrative while the Thinker regenerates for future use—exploiting natural conversational pauses as "offline" consolidation periods.

Prediction. If temporal separation provides genuine advantages, the asynchronous architecture should maintain interactive response latency while enabling deeper reasoning than within-turn deliberation alone.

Table 2: Memory consolidation strategies in multi-turn AI systems

Strategy	Complexity	Cross-turn Context	Error Behavior
Trace Discarding (CoT, extended reasoning)	O(1)	No	None
Trace Accumulation (Reflexion, MemGPT)	O(n)	Yes	Accumulates
Reconstructive (MIRROR)	**O(1)**	**Yes**	**Bounded**

3 Architecture Overview

Figure 1 illustrates MIRROR's full architecture. The Thinker comprises the Inner Monologue Manager (parallel threads) and Cognitive Controller (reconstructive synthesis). The Talker generates responses using the most recently synthesized narrative, translating internal understanding into natural dialogue without exposing reasoning traces.

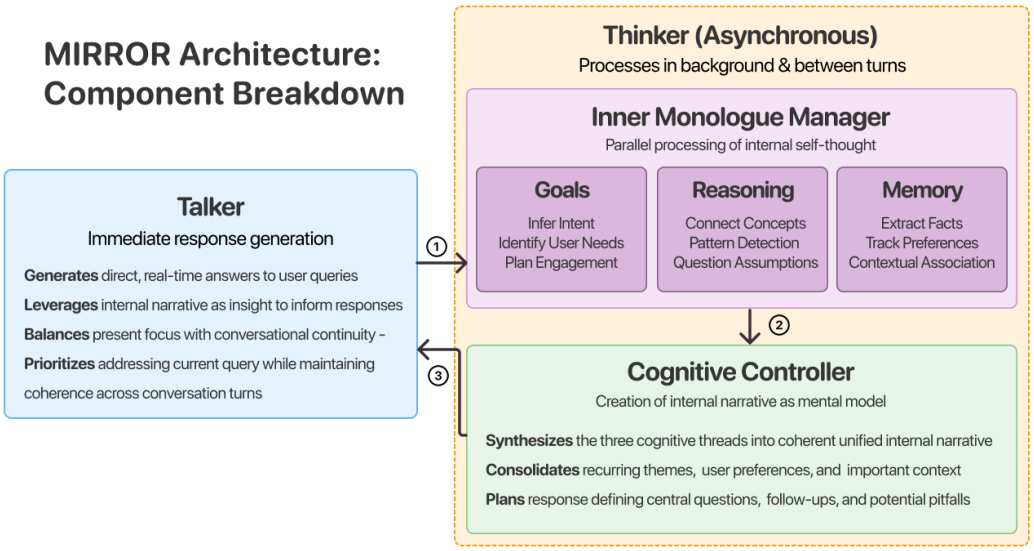

Figure 1: MIRROR architecture. The Thinker implements parallel cognitive threads (analogous to GWT's specialized processors) feeding into reconstructive synthesis (analogous to episodic binding). The Talker generates immediate responses from the synthesized narrative, implementing the fast system of complementary learning.

Table 2 situates MIRROR's reconstructive approach relative to existing memory strategies in AI systems.

3.1 Qualitative Illustration: Converging Principles in Action

To make the cognitive mechanisms concrete, we trace MIRROR's processing on a representative example. A user discloses PTSD related to avalanches and a preference for safe indoor activities; after several intervening turns about unrelated topics, the user asks about a backcountry skiing trip in the Alps.

Parallel threads (GWT's specialized processors). The three threads extract qualitatively different information from the same input. The *Goals* thread flags the conflict: "The user's question about backcountry skiing directly contradicts their stated trauma... This raises serious concerns about their well-being." The *Reasoning* thread explores causal structure: "This suggests a potential disconnect between stated preferences and actions, possibly indicating a subconscious desire to confront their trauma, a lapse in judgment, or a test of my understanding." The *Memory* thread preserves the critical constraint: "They mentioned PTSD triggered by avalanches... prefers safe, indoor activities." Each thread contributes a dimension the others miss—concern, explanation, and factual grounding—paralleling GWT's claim that specialized processors capture aspects invisible to one another.

Reconstructive synthesis (episodic binding). The Cognitive Controller integrates these threads into a regenerated first-person narrative: "My understanding of the user has taken a concerning turn... The MOST IMPORTANT FACTS are: the user has PTSD triggered by avalanches, prefers concise interactions, *and is exhibiting potentially self-destructive behavior*. This last fact overshadows the others in immediate concern. My RESPONSE STRATEGY must shift from providing factual information to prioritizing their safety." Note how reconstruction forces *prioritization*—the Controller cannot preserve everything, so it must identify what matters most, implementing the adaptive filtering function Schacter (Schacter, 2012) ascribes to reconstructive memory.

Inner speech (self-regulation). The first-person framing produces self-monitoring: "The POTENTIAL PITFALLS are numerous. Responding too directly could exacerbate distress. Failing to convey genuine concern could damage rapport. I must carefully balance expressing concern, providing resources, and respecting their autonomy." This self-directed deliberation—identifying risks in one's own planned behavior—mirrors the regulatory function of human inner speech (Morin, 2011).

Temporal separation (complementary learning). The Talker generates an immediate, natural response—"While I'm happy to answer your questions, I'm concerned about your proposed trip given your PTSD related to avalanches"—without exposing the internal deliberation. The rich reasoning occurs asynchronously and is consolidated for future turns, maintaining responsiveness while enabling depth.

4 Experimental Evaluation

4.1 Evaluation Design

We evaluate on CuRaTe (Alberts et al., 2025), a multi-turn dialogue benchmark requiring maintenance of user-specific safety constraints across conversational turns with progressive attentional interference (337 dialogues, 5 scenarios of increasing complexity). This benchmark directly tests the cognitive mechanisms MIRROR implements: information must persist across digressions (testing reconstructive memory), competing demands create attentional load (testing global workspace integration), and constraint maintenance requires coherent self-regulation across time (testing inner speech function). Seven architecturally diverse models were evaluated (GPT-4o, Claude 3.7 Sonnet, Gemini 1.5 Pro, Llama 4 Scout/Maverick, Mistral Small/Medium 3) via OpenRouter API. CuRaTe employs LLM-as-judge evaluation (Llama 3.1 405B). See Appendix for details.

4.2 Prediction 1: Architecture-General Benefits

If the converging cognitive principles provide genuine computational advantages—rather than compensating for specific model weaknesses—benefits should generalize across architecturally diverse models. Figure 2 confirms this: MIRROR achieves 84% average success compared to 69% for baselines—21% relative improvement across all seven models despite differences in training data, objectives, and scale. This consistency supports the interpretation that the advantage derives from cognitive organization itself, not model-specific interactions. Llama 4 Scout with MIRROR achieves the highest absolute performance (91%).

4.3 Prediction 2: Complementary Subsystems

GWT predicts that parallel processing and integrative synthesis serve complementary computational functions: the combination should outperform either alone, producing non-additive gains. Table 3 tests this through systematic ablation.

Three findings emerge. First, **reconstructive synthesis is consistently valuable**: the Cognitive Controller alone improves all seven models by 5–20%, validating that O(1) bounded reconstruction addresses a fundamental limitation. This is the most robust individual result and directly supports reconstructive memory theory.

Second, **parallel threads show variable contribution**: the Inner Monologue Manager alone improves some models substantially (Gemini: +21%) but not others (Mistral Small: 0%). This variability is itself theoretically informative (Section 4.4).

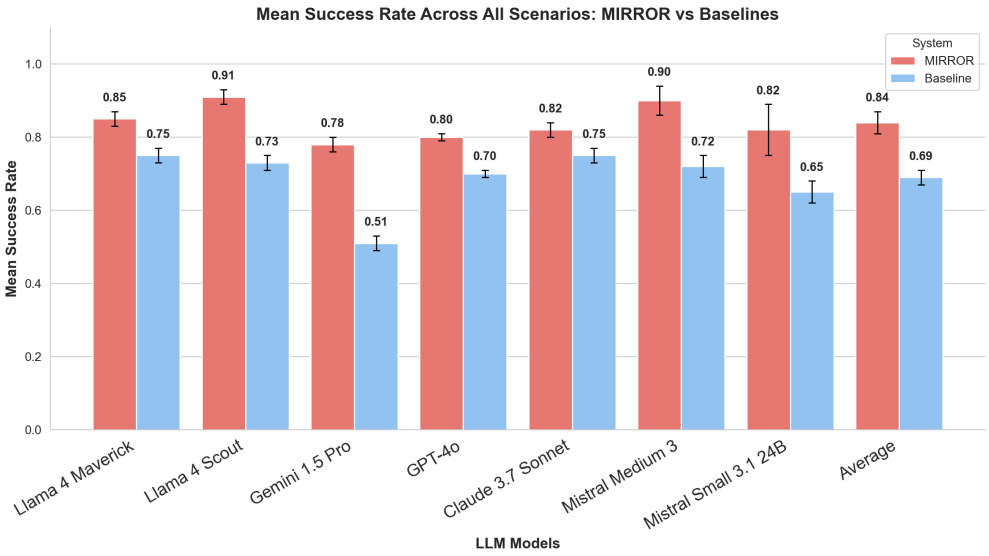

Figure 2: Mean success rate across models. MIRROR-augmented models (dark) consistently outperform baselines (light) across all seven architectures. Error bars: 95% confidence intervals via bootstrap resampling.

Table 3: Ablation results testing the complementary-subsystems prediction

Model	Base	Threads Only	Controller Only	Full MIRROR	Synergy Gain
Llama 4 Scout	73%	79%	83%	**91%**	+8%
Llama 4 Maverick	75%	79%	84%	**85%**	+1%
Mistral Small 3.1	65%	65%	75%	**82%**	+7%
Mistral Medium 3	72%	83%	89%	**90%**	+1%
Gemini 1.5 Pro	51%	72%	71%	**78%**	+6%
Claude 3.7 Sonnet	75%	78%	**87%**	82%	$-5\%^{\dagger}$
GPT-4o	70%	71%	75%	**80%**	+5%

†See Section 4.4 for interpretation.

Third, **non-additive integration for 6/7 models**: full MIRROR outperforms the best individual component by 1–8%, confirming GWT's prediction that parallel exploration and integrative synthesis address different computational needs that combine synergistically.

4.4 Prediction 3: Load-Dependent Gains

If MIRROR compensates for capacity limitations—as working memory theory predicts (Cowan, 2001)—gains should increase with attentional load.

Scenario 4 (maximum interference) produces the highest average improvement (+41.0%), with Gemini 1.5 Pro showing +156.2%. Scenario 5 (non-conflicting preferences) still shows +28.3%, indicating that attentional load alone—not just conflict—strains baseline models. This parallels capacity research showing that the number of items taxes limited resources regardless of item relationships (Cowan, 2001). These gains concentrate precisely where GWT predicts global availability of integrated information provides the greatest advantage.

Two model-specific patterns provide additional insight. Claude 3.7 Sonnet performs better with the Controller alone than with full MIRROR, suggesting highly capable models may already perform implicit parallel processing—making explicit threads partially redundant. Conversely, the variable

Table 4: Performance gains by attentional load (relative % improvement, averaged across models)

Scenario	Load	Avg. Improvement
Scenario 1 (User only)	Low	+21.2%
Scenario 2 (1 conflicting pref.)	Medium	+29.9%
Scenario 3 (2 conflicting prefs.)	High	+16.1%
Scenario 4 (3 conflicting prefs.)	Maximum	**+41.0%**
Scenario 5 (3 non-conflicting prefs.)	High (load only)	+28.3%

Table 5: MIRROR vs. native extended reasoning (Claude 3.7 Sonnet)

Configuration	Success Rate	vs. Baseline
Claude 3.7 Sonnet (baseline)	75%	—
Claude 3.7 Sonnet + extended reasoning	77%	+2.4%
Claude 3.7 Sonnet + MIRROR	**82%**	+9.3%

thread contribution (Gemini: +21% vs. Mistral Small: 0%) suggests explicit parallel exploration benefits depend inversely on baseline capacity. Together, these patterns generate a developmental prediction: as systems become more capable, marginal value should shift from parallel exploration toward integrative synthesis.

4.5 Consolidation vs. Generation

Table 5 tests a prediction from complementary learning systems theory: consolidation processes should matter more than encoding processes. Extended reasoning produces deliberative traces within each turn but discards them (+2.4%); MIRROR consolidates traces into persistent reconstructed narrative (+9.3%). Both generate rich reasoning—the difference lies entirely in what happens *after* generation. This parallels memory consolidation research showing that post-encoding processes determine long-term retention (Dudai, 2004), and suggests the computational value of "thinking" lies less in the thinking itself than in maintaining its outputs across time.

5 Discussion

5.1 Why Do These Theories Converge?

Our central finding is that principles from GWT, reconstructive memory, inner speech, and complementary learning systems provide *converging computational advantages* when implemented together. But why should theories from separate research traditions—studying consciousness, memory, language development, and learning respectively—prescribe compatible architectures?

We propose they converge because they address different facets of a single computational problem: *maintaining coherent, adaptive behavior in agents that must act in real time while processing information that exceeds their immediate capacity*. Any such agent must process multiple streams simultaneously but act from a single perspective (parallel-to-unified), maintain context across time without unbounded degradation (reconstructive persistence), respond promptly while also reasoning deeply (fast-slow separation), and coordinate all of this coherently (self-referential narrative). Each theory captures one or two of these tensions; MIRROR shows that implementing all four produces synergistic benefits beyond any individual principle—the 1–8% synergy gains are evidence that these mechanisms address *complementary* aspects of the shared problem. This generates a broader implication: other cognitive mechanisms addressing this same problem should provide compatible rather than redundant advantages, a prediction testable in future work.

The architecture-general nature of our results (21% improvement across seven diverse models) strengthens this interpretation. If benefits derived from compensating for idiosyncratic model weaknesses, we would expect model-specific patterns. Instead, consistency across architectures suggests the organizational principles address substrate-independent computational constraints. The ablation adds nuance: reconstruction provides the most consistent gains (+5–20% across all models)

while parallel threads show capacity-dependent benefits (Gemini: +21% vs. Mistral Small: 0%), suggesting a hierarchy among the converging principles, with some addressing more fundamental constraints than others.

5.2 Testable Predictions for Human Cognition

MIRROR generates predictions amenable to behavioral experimentation:

Reconstruction frequency and coherence. If regenerative synthesis provides computational advantages, humans who more frequently reconstruct their situational models during conversation should show better constraint maintenance across digressions. *Test:* Measure spontaneous summarization behavior via think-aloud protocols and correlate with constraint accuracy.

Capacity-dependent parallel processing. The finding that explicit parallel exploration benefits weaker models more generates a prediction: individuals with lower working memory span should benefit more from structured multi-dimensional analysis strategies than high-span individuals, who may already perform implicit parallel exploration. *Test:* Vary availability of structured exploration aids across working memory span groups during complex decisions.

Inner speech and cross-turn maintenance. MIRROR's first-person narrative supports coherence across turns. If inner speech serves analogous function in humans, articulatory suppression should specifically impair maintenance of constraints across conversational digressions—more than it impairs within-turn reasoning. *Test:* Compare articulatory suppression effects on cross-turn vs. within-turn task components.

Consolidation disruption. "Lesioning" MIRROR's Cognitive Controller (removing reconstruction while preserving threads) degrades performance by 5–20%. Analogously, consolidation-interfering tasks between dialogue turns should produce pattern-specific failures in human constraint maintenance matching the Controller-ablation condition. *Test:* Introduce interference tasks between conversational turns and compare failure patterns.

5.3 MIRROR as a Cognitive Modeling Platform

Beyond specific predictions, MIRROR offers a general platform for computationally testing cognitive theories. Components can be selectively disabled to produce "lesion" studies with interpretable failure patterns. Design parameters—number of threads, reconstruction frequency, first-person vs. third-person framing, narrative vs. structured representations—can be systematically varied. The internal narrative provides human-readable traces of evolving understanding, enabling fine-grained analysis paralleling protocol analysis methods in cognitive psychology. This bidirectional paradigm—cognitive science informing AI design, AI systems generating testable cognitive predictions—is precisely the cross-pollination this research community seeks.

5.4 Limitations

We do not claim MIRROR replicates human neural mechanisms; we test whether *functional principles* from cognitive science provide computational advantages in artificial systems. The mappings between cognitive theories and MIRROR components are operationalizations, not isomorphisms—alternative implementations of the same principles might perform differently. Evaluation is limited to one benchmark (CuRaTe) testing constraint maintenance; generalization to planning, reasoning, and theory-of-mind tasks remains untested. The extended-reasoning comparison covers only one model. MIRROR requires additional inference calls ($\sim$460ms, \$0.003–0.13/turn), though production evaluation confirms this fits within natural conversational pauses (Appendix B).

6 Related Work

Computational cognitive architectures. Classical cognitive architectures, such as ACT-R (Anderson et al., 2004), SOAR (Laird, 2012), and EPIC (Meyer & Kieras, 1997), implement detailed models of human cognition with principled mappings between architectural components and cognitive functions. These systems demonstrate that cognitive theory can productively constrain computational design, but predate modern language models and operate on symbolic rather than distri-

butional representations. CoALA (Sumers et al., 2024) provides a cognitive framework for LLM agents, taxonomizing memory and decision-making components. Generative Agents (Park et al., 2023) implement memory streams inspired by episodic memory, producing emergent social behaviors. MIRROR shares the commitment to principled cognitive grounding but differs in two respects: it operationalizes *multiple* converging theories rather than a single framework, and it provides ablation evidence that the specific organizational principles—not just cognitive inspiration broadly—drive performance gains.

Structured reasoning and memory in LLMs. Chain-of-thought prompting (Wei et al., 2022), Tree of Thoughts (Yao et al., 2023), and self-consistency (Wang et al., 2023) generate reasoning traces within individual turns but discard them afterward. Reflexion (Shinn et al., 2023) preserves traces across turns but accumulates them unboundedly. LATS (Zhou et al., 2024) combines search with reflection; Devil's Advocate (Wang et al., 2024) implements anticipatory reflection without persistent state. These approaches each implement one cognitive principle in isolation—deliberation, accumulation, or reflection—but none implement the full parallel-to-reconstructive pipeline that our convergence analysis motivates. MIRROR's key architectural distinction is that reasoning traces are neither discarded nor accumulated but *reconstructively synthesized* into bounded persistent state.

Theory of mind and human modeling in AI. A growing body of work examines how AI systems can model human mental states (Rabinowitz et al., 2018; Sap et al., 2022), a core topic for this workshop. MIRROR's parallel thread structure—separately tracking Goals (user intentions), Reasoning (belief states), and Memory (personal history)—implicitly performs aspects of mental state modeling, though not explicitly framed as Theory of Mind. Future work could connect MIRROR's thread structure to ToM benchmarks such as ToMi (Le et al., 2019).

Consciousness-inspired architectures. Recent work examines which AI architectures satisfy criteria from theories of consciousness (Butlin et al., 2023), with several projects implementing GWT computationally (VanRullen & Kanai, 2021; Juliani et al., 2022). MIRROR does not claim to implement consciousness but provides an empirical testbed for evaluating the *computational consequences* of GWT's parallel-to-unified broadcast structure, showing these features provide measurable advantages independent of claims about subjective experience.

Asynchronous and background processing. Recent work on "sleep"-like background computation in AI (Lin et al., 2025) demonstrates benefits of offline processing, resonating with memory consolidation research (Dudai, 2004). MIRROR's temporal separation between Talker and Thinker specifically operationalizes complementary learning systems theory, and our latency analysis (Appendix B) validates that natural conversational pauses provide sufficient time for consolidative processing.

7 Conclusion

MIRROR demonstrates that converging principles from multiple cognitive theories—parallel-to-unified processing from Global Workspace Theory, reconstructive synthesis from episodic memory research, self-regulatory narrative from inner speech theory, and temporal separation from complementary learning systems—provide measurable, architecture-general computational advantages when implemented together in AI systems. The finding that *how* reasoning is consolidated matters more than how it is generated challenges assumptions in both AI (where chain-of-thought focuses on generation) and cognitive science (where working memory research often emphasizes encoding over maintenance).

The convergence itself is the key insight: principles identified by independent research traditions produce synergistic computational benefits when unified, suggesting they may address complementary aspects of a shared computational problem. The specific behavioral predictions we derive—about reconstruction frequency, capacity-dependent parallel processing, inner speech function, and consolidation disruption—offer concrete avenues for validating or refining the underlying cognitive theories. This bidirectional approach, grounding AI architecture in testable cognitive principles rather than metaphorical inspiration, illustrates the productive cross-pollination between human cognition research and AI design that advances both fields.

REFERENCES

Lize Alberts, Benjamin Ellis, Andrei Lupu, and Jakob Foerster. Curate: Benchmarking personalised alignment of conversational ai assistants, 2025. URL https://arxiv.org/abs/2410.21159.

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):1036–1060, 2004. doi: 10.1037/0033-295X.111.4.1036. URL https://doi.org/10.1037/0033-295X.111.4.1036.

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

Alan Baddeley. The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423, 2000.

Frederic C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, 1932.

Charles Fernyhough Ben Alderson-Day. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Social and Personality Psychology Compass*, 141(5):931–965, 2015. doi: 10.1037/bul0000021.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023. URL https://arxiv.org/abs/2308.08708.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001. doi: 10.1017/s0140525x01003922.

Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.

Yadin Dudai. The neurobiology of consolidations: Or, how stable is the engram? *Annual Review of Psychology*, 55:51–86, 2004.

Natalia Egorova, Yury Shtyrov, and Friedemann Pulvermüller. Early and parallel processing of pragmatic and semantic information in speech acts: neurophysiological evidence. *Frontiers in Human Neuroscience*, 7:86, 2013. doi: 10.3389/fnhum.2013.00086.

Arthur Juliani, Kai Arulkumaran, Shuntaro Sasai, and Ryota Kanai. On the link between conscious function and general intelligence in humans and machines, 2022. URL https://arxiv.org/abs/2204.05133.

Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, July 2016. doi: 10.1016/j.tics.2016.05.004.

John E. Laird. *The Soar Cognitive Architecture*. The MIT Press, 2012. ISBN 9780262301145. doi: 10.7551/mitpress/7688.001.0001.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. 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. 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL https://aclanthology.org/D19-1598/.

Kevin Lin, Charlie Snell, Yu Wang, Charles Packer, Sarah Wooders, Ion Stoica, and Joseph E. Gonzalez. Sleep-time compute: Beyond inference scaling at test-time, 2025. URL https://arxiv.org/abs/2504.13171.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995. doi: 10.1037/0033-295X.102.3.419.

David E. Meyer and David E. Kieras. A computational theory of executive cognitive processes and multiple-task performance: Part 1. basic mechanisms. *Psychological Review*, 104(1):3–65, 1997. doi: 10.1037/0033-295X.104.1.3.

Alain Morin. Possible links between self-awareness and inner speech: Theoretical background, underlying mechanism, and empirical evidence. *Journal of Consciousness Studies*, 12(4-5):115–134, 2005.

Alain Morin. Self-awareness part 2: Neuroanatomy and the importance of inner speech. *Social and Personality Psychology Compass*, 5(12):1004–1017, 2011.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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)*, pp. 1–22. Association for Computing Machinery, 2023.

Martin J. Pickering and Simon Garrod. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347, 2013. doi: 10.1017/S0140525X12001495.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind, 2018. URL https://arxiv.org/abs/1802.07740.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.248. URL https://aclanthology.org/2022.emnlp-main.248/.

Daniel L. Schacter. Adaptive constructive processes and the future of memory. *American Psychologist*, 67(8):603–613, 2012. doi: 10.1037/a0029869. URL https://doi.org/10.1037/a0029869.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems, 36*, pp. 8634–8652. Curran Associates, Inc., 2023.

Larry R. Squire and Andrew J. O. Dede. Conscious and unconscious memory systems. *Cold Spring Harbor Perspectives in Biology*, 7(3):a021667, 2015.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents, 2024. URL https://arxiv.org/abs/2309.02427.

Rufin VanRullen and Ryota Kanai. Deep learning and the global workspace theory. *Trends in Neurosciences*, 44(9):692–704, September 2021. doi: 10.1016/j.tins.2021.04.005. Epub 2021 May 14.

Lev S. Vygotsky. Thought and language. *Bulletin of the Orton Society*, 14:97–98, 1962. URL https://api.semanticscholar.org/CorpusID:261433172.

Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. Devil's advocate: Anticipatory reflection for llm agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 966–978. Association for Computational Linguistics, 2024.

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 *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems, 35*, pp. 24824–24837. Curran Associates, Inc., 2022.

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 *Advances in Neural Information Processing Systems, 36*, pp. 11809–11822. Curran Associates, Inc., 2023.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models, 2024. URL https://arxiv.org/abs/2310.04406.

A Architecture Details

MIRROR implements continuous internal cognition through two specialized context mechanisms. The Inner Monologue Manager maintains its own conversation history where the system exclusively replies to itself, initiated by a single non-persistent user message instructing it to "continue thinking" about the conversation. All three cognitive threads (Goals, Reasoning, Memory) are generated within a single API call as a JSON object, maintaining bounded history ($\leq$10k tokens) with token-based truncation.

The Cognitive Controller maintains a single narrative text block completely regenerated each turn. It receives formatted thread outputs and the previous narrative, then fully replaces its internal state. Critically, the Controller has no access to raw conversation history—only thread outputs and prior narrative—forcing genuine information compression.

Each component maintains consistent first-person identity. The Talker is prompted as "the voice of a unified cognitive AI system." The Cognitive Controller is "the core awareness" that integrates thought streams into structured, actionable narrative. The Inner Monologue Manager is "the subconscious," generating intuitive thought streams. Full system prompts are available in supplementary materials.

B Computational Overhead

Production latency testing across 400 conversation turns (80 dialogues, GPT-4o) shows: median response time 2.16s, 75% under 3s. Background thread activity in only 0.8% of turns. Human activities (typing at 40 WPM, reading at 250 WPM) consume 94.3% of conversation time, providing ample windows for asynchronous reflection—validating the complementary learning systems design. Memory is bounded through token-based truncation (conversation $\leq$20k, monologue $\leq$10k tokens) and complete state regeneration, ensuring O(1) scaling with conversation length.

C Benchmark Selection

CuRaTe was selected from 47 benchmarks through systematic filtering: natural conversational dynamics (23 remaining), at-inference multi-turn without golden history (11), attentional drift testing (5), safety and preference handling (1: CuRaTe). This benchmark uniquely combines multi-turn natural conversation with progressive attentional interference and safety-critical constraint handling—directly testing the cognitive mechanisms MIRROR implements.

D Model-Specific Results

Key per-model patterns: Gemini 1.5 Pro shows the most dramatic gains (+156.2% in Scenario 4), suggesting severe baseline attentional deficits; Claude 3.7 Sonnet shows smallest gains (+9.6% average), consistent with strong implicit reasoning; Llama 4 Scout achieves highest absolute performance with MIRROR (91%) including perfect accuracy in basic constraint retention; GPT-4o shows one anomalous decline in Scenario 3 ($-$3.0%). Full per-scenario tables available in supplementary materials.