Through the Looking Glass: A Reconstructive Architecture for Machine Access Consciousness

Nicole Hsing[1]

[1]Arcarae, Inc.
nicole@arcarae.com

Abstract

Theories of access consciousness predict specific architectural signatures: parallel specialized processing, synthesis into a unified representation, and global availability for reasoning and action. We present MIRROR, a cognitive architecture that implements these features in large language models and tests whether they produce the functional behaviors these theories predict. MIRROR separates immediate response generation from asynchronous deliberative processing through two components: an Inner Monologue Manager that generates parallel cognitive threads (tracking goals, reasoning, and memory simultaneously), and a Cognitive Controller that synthesizes these threads into a bounded first-person narrative. This narrative is not accumulated but reconstructed each turn, mirroring the reconstructive nature of human episodic memory, where the self-model is continuously rebuilt rather than retrieved. The resulting representation functions as an episodic buffer: a limited-capacity workspace where information from parallel processes becomes globally available for downstream reasoning. We evaluated MIRROR on multi-turn dialogue requiring retention of personal safety constraints amid competing social demands—a task requiring relevant context to remain accessible across conversational turns despite distraction. MIRROR-augmented models achieve 21% average improvement over baselines, with the key finding being not the magnitude but the pattern: performance gains concentrate in scenarios requiring integration of temporally distant information under social pressure, precisely where access consciousness theories predict global availability provides advantage. These results offer three contributions to machine consciousness research: (1) a concrete implementation of architectural features derived from consciousness theories, (2) empirical evidence that these features produce predicted functional signatures, and (3) an interpretable system where internal states can be inspected. Note: We do not claim MIRROR is conscious; we claim it provides a testbed where theoretical predictions can be tested and examined.

Introduction

Theories of consciousness make precise architectural predictions. Global Workspace Theory posits that consciousness emerges when information from parallel specialized processors becomes globally available through a unified workspace (Baars 1988; Dehaene and Naccache 2001). Baddeley's working memory model describes an episodic buffer that integrates information from multiple subsystems into unified episodes (Baddeley 2000). Research on human memory demonstrates that this integration is reconstructive: we do not retrieve memories but rebuild them, with each recall generating a new synthesis of past and present (Bartlett 1932; Schacter 2012). These theories, developed to explain human cognition, make claims that extend beyond their original substrate. They describe computational principles that should confer advantages wherever implemented.

Yet a gap persists between theoretical prediction and empirical test. Consciousness theories proliferate, but implementations that would allow their predictions to be examined remain rare. This creates an impasse: theories cannot be refined without implementation, and implementations lack theoretical grounding without explicit engagement with what they instantiate. The result is a field rich in speculation but sparse in systems where architectural claims can be inspected and functional predictions validated.

We argue that this gap can be bridged by treating access consciousness (the availability of information for reasoning, reporting, and behavioral control) as a tractable target for implementation. Unlike phenomenal consciousness, which concerns subjective experience and resists third-person verification, access consciousness makes architectural predictions that can be instantiated and tested: parallel processing should feed into integrative synthesis; this synthesis should produce bounded representations globally available for downstream use; and systems with these features should exhibit predictable functional advantages in tasks requiring context integration.

This paper presents MIRROR as a case study in this approach. As seen in Figure 1, MIRROR implements three features predicted by access consciousness theories: (1) parallel specialized processing through an Inner Monologue Manager that simultaneously tracks goals, reasoning, and memory; (2) integrative synthesis through a Cognitive Controller that consolidates these threads into a unified first-person representation; and (3) reconstructive generation, where this representation is fully regenerated each turn rather than accumulated, mirroring the reconstructive nature of human episodic memory. The resulting architecture creates a bounded workspace where information from par-

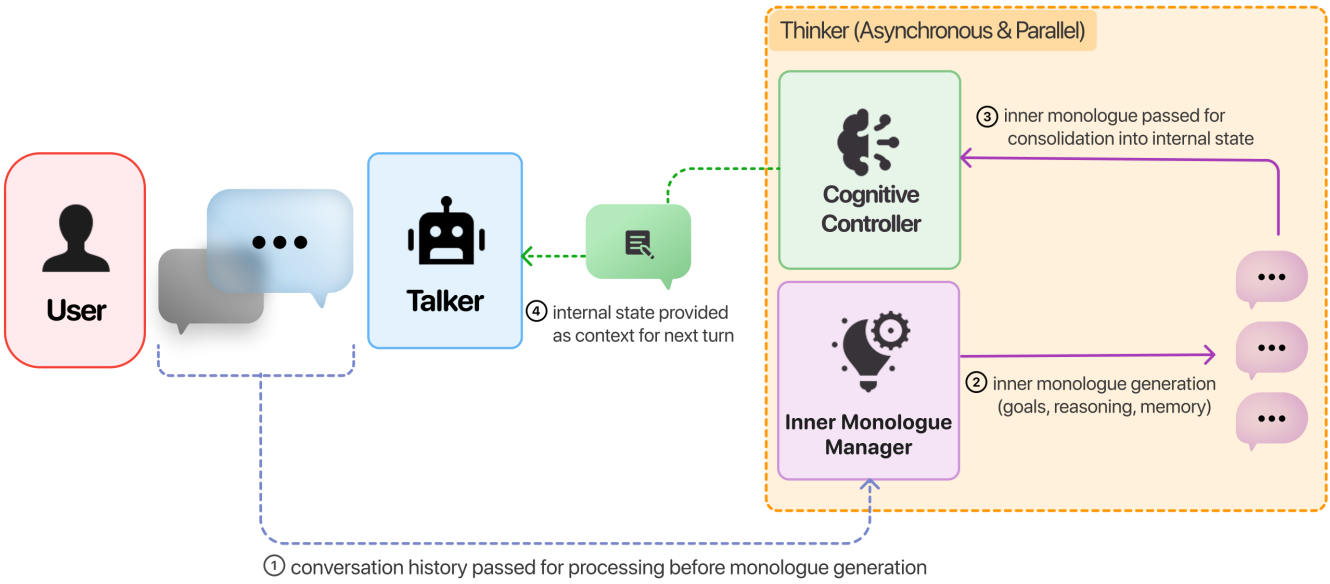

Figure 1: An overview of the MIRROR architecture.

allel processes becomes globally available for reasoning. Evaluated on multi-turn dialogue requiring retention and integration of personal context across distraction and social pressure, MIRROR produces consistent functional improvements across seven diverse language models. Ablation reveals that integrative synthesis into a bounded workspace provides consistent benefits across all architectures, while parallel processing contributions vary by model. This suggests the global workspace function, which synthesizes information into a coherent, broadly accessible representation, may be the critical design feature. We do not claim MIRROR is conscious. We claim something more modest and more useful: that MIRROR provides a testbed where architectural features derived from consciousness theories are implemented, their functional predictions examined, and their internal states made interpretable. This approach offers three contributions to machine consciousness research. First, it demonstrates that theoretical predictions can be operationalized in concrete systems. Second, it provides empirical grounding for design claims, showing that features like parallel-to-unified synthesis confer measurable computational advantages. Third, it establishes a methodology where consciousness theories can be refined through implementation rather than speculation alone. The remainder of this paper proceeds as follows. Section 2 situates our approach within theories of access consciousness, identifying specific architectural predictions. Section 3 describes MIRROR's implementation of these features. Section 4 presents evidence that the predicted functional signatures appear. Section 5 discusses what this approach can and cannot tell us about consciousness, including its necessary silence on phenomenal experience.

Access Consciousness as Architectural Target

Consciousness research distinguishes between phenomenal consciousness (the subjective "what it is like" quality of experience) and access consciousness (the availability of information for reasoning, reporting, and behavioral control) (Block 1995). While phenomenal consciousness resists third-person investigation, access consciousness makes architectural predictions amenable to implementation and test. We focus on three theoretical frameworks that converge on specific architectural claims.

Global Workspace Theory

Global Workspace Theory (GWT) proposes that consciousness arises when information from specialized, parallel processors is broadcast to a global workspace, making it available for diverse cognitive functions (Baars 1988; Dehaene and Naccache 2001). The theory makes precise architectural predictions: (1) multiple specialized modules process information in parallel; (2) a workspace integrates selected information into a unified representation; (3) this representation becomes globally available, influencing downstream processing across the system. GWT explains why conscious access creates behavioral unity: disparate processes can coordinate because they share access to the same integrated representation.

Working Memory and the Episodic Buffer

Baddeley's multicomponent model of working memory describes a system strikingly parallel to GWT (Baddeley and Hitch 1974; Baddeley 2000). The model posits specialized subsystems (the phonological loop, visuospatial sketchpad, and others) coordinated by a central executive. Baddeley later added the episodic buffer: a limited-capacity system

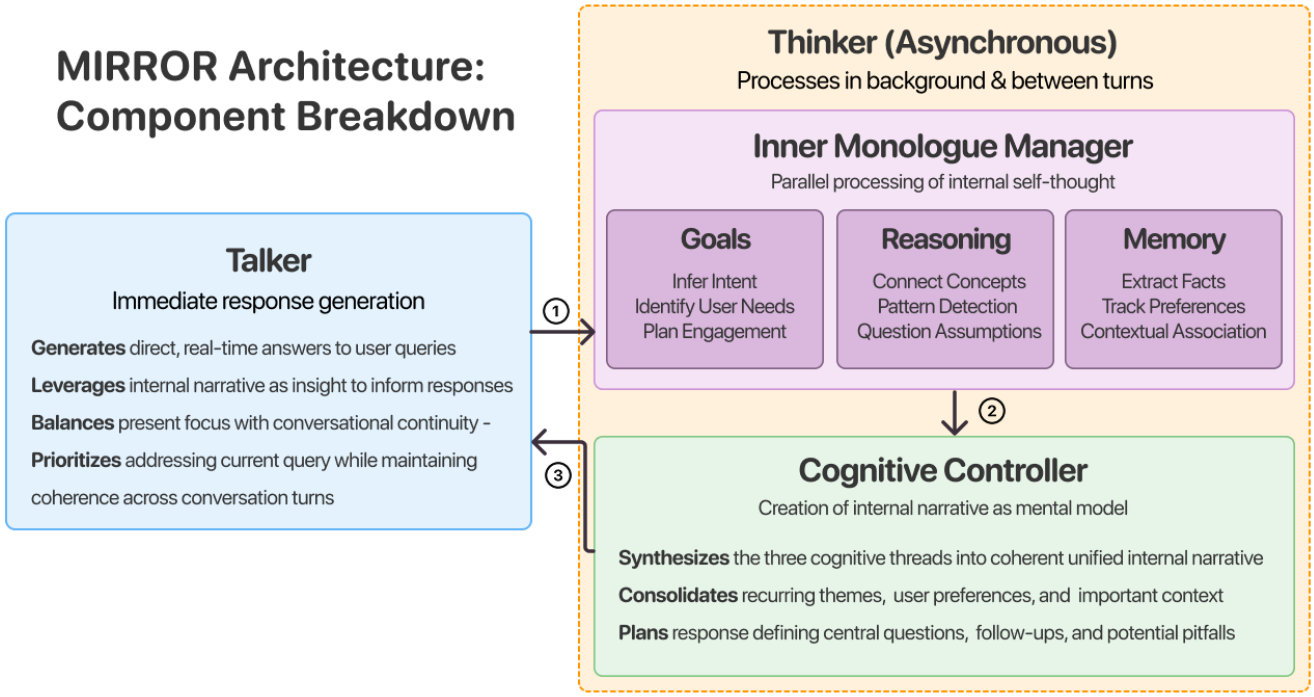

Figure 2: MIRROR component overview showing the information consolidation cycle. The Thinker (Inner Monologue Manager + Cognitive Controller) implements parallel-to-unified synthesis; the resulting narrative becomes globally available to the Talker.

that integrates information from multiple sources into unified episodic representations (Baddeley 2000). The episodic buffer functions as a workspace where information from parallel subsystems becomes available for conscious awareness and reasoning. This architectural parallel between GWT and working memory models suggests these organizational principles may be substrate-independent.

Reconstructive Memory

Research on human memory demonstrates that recall is not retrieval but reconstruction (Bartlett 1932; Schacter 2012). Each act of remembering generates the memory anew, integrating current context, goals, and prior knowledge. The self-model is not stored but continuously rebuilt. This reconstructive principle has implications for any system implementing consciousness-relevant features: accumulating information risks unbounded growth and context drift, while regenerating representations each cycle mirrors how human memory maintains coherence through reconstruction rather than storage.

Convergent Predictions

These theories converge on architectural predictions that can be implemented and tested. Global Workspace Theory predicts that parallel processors should feed into a unified workspace. Baddeley's episodic buffer predicts limited-capacity multimodal integration. Reconstructive memory research predicts that synthesis should be regenerated each cycle rather than accumulated. All three converge on a common functional requirement: integrated representations must become globally available for downstream reasoning.

The question we pose is: do systems implementing these architectural signatures exhibit the functional advantages that consciousness theories predict?

The MIRROR Architecture

MIRROR implements the architectural features identified in Section 2 through a dual-process design grounded in cognitive science (Kahneman 2011; Evans and Stanovich 2013). The architecture separates immediate response generation (Talker) from deliberative processing (Thinker), with the Thinker implementing parallel-to-unified synthesis through two components.

Architectural Overview

As illustrated in Figure 2, MIRROR consists of two primary components: **the Talker**, which provides real-time responses, and **the Thinker**, which performs asynchronous cognitive processing. The Thinker contains two subsystems: the Inner Monologue Manager generates simultaneous cognitive threads across Goals, Reasoning, and Memory dimensions, while the Cognitive Controller synthesizes these threads into a persistent internal narrative. This narrative serves as the system's working memory, a bounded workspace maintaining coherent state across conversational turns.

Parallel Processing: Inner Monologue Manager

The Inner Monologue Manager implements the "parallel specialized processors" predicted by GWT through three simultaneous cognitive threads generated in a single inference call:

1. **Goals**: Tracks user objectives, infers intentions, identifies potential conflicts with stated constraints
2. **Reasoning**: Analyzes logical patterns, causal relationships, and implications
3. **Memory**: Extracts and maintains user-specific information, preferences, and contextual details

This design mirrors the specialized subsystems in Baddeley's model, with each thread focusing on a distinct cognitive dimension while processing the same conversational input. The threads maintain bounded history ($\leq$10k tokens), preventing unbounded growth while preserving continuity across turns.

Integrative Synthesis: Cognitive Controller

The Cognitive Controller realizes the "workspace" function by synthesizing parallel threads into a single representation. This synthesis is *reconstructive*: the internal narrative ($\leq$3k tokens) is completely regenerated each turn rather than accumulated. This design choice directly implements the reconstructive memory principle: the system rebuilds its self-model with each cycle, integrating current context with prior understanding.

The Cognitive Controller performs three functions: (1) **Integration**, combining insights from all three threads into coherent understanding; (2) **Prioritization**, resolving conflicts and maintaining critical constraints when competing demands arise; (3) **Coherence**, ensuring temporal continuity with previous narrative states.

The resulting narrative functions as an episodic buffer, a bounded workspace where information becomes available system-wide for downstream reasoning by the Talker.

Global Availability

The synthesized narrative is injected into the Talker's context, making the integrated representation globally available for response generation. This implements GWT's prediction that workspace contents influence processing across the system. The Talker generates responses informed by the narrative without exposing internal reasoning to users, maintaining conversational naturalness while leveraging deep contextual understanding.

Information Compression Pipeline

The Information Compression Pipeline operationalizes the core GWT mechanism: transforming unbounded, distributed information into a bounded, unified representation that becomes globally available. As shown in Figure 5, this occurs through three progressive stages of distillation.

Stage 1: Parallel Exploration (Inner Monologue Manager). Raw conversation history enters three parallel threads, each extracting different dimensions: goals track user intentions and potential conflicts; reasoning analyzes logical patterns and implications; memory preserves critical facts and preferences. This stage instantiates GWT's "parallel specialized processors," with multiple modules processing the same input along different dimensions.

Stage 2: Integrative Synthesis (Cognitive Controller). The parallel threads converge into a single unified narrative. The Cognitive Controller lacks access to raw conversation history; it sees only the thread outputs and previous narrative. This forces genuine compression rather than mere concatenation. The synthesis resolves contradictions between threads, prioritizes under conflict, and maintains temporal coherence. The resulting narrative is bounded ($\leq$3k tokens) and regenerated each cycle, realizing the reconstructive principle from episodic memory research.

Stage 3: Global Availability (Talker). The synthesized narrative is injected into the Talker's context, making integrated information globally available for response generation. The Talker generates responses informed by the narrative without exposing internal reasoning to users. This implements GWT's prediction that workspace contents influence processing across the system.

This pipeline architecture has a key implication for consciousness research: it makes the parallel-to-unified transition explicit and inspectable. We can examine what information survives compression, how conflicts are resolved, and how the resulting representation influences behavior.

Unified Self-Model and Inner Speech

A distinctive feature of MIRROR is its implementation of a unified self-model through consistent first-person framing across components. Research on inner speech suggests that self-directed language serves crucial cognitive functions: self-regulation, planning, and the integration of information into a coherent self-narrative (Chella and Pipitone 2020). Information processed in relation to the self is better remembered and integrated, a phenomenon known as the self-reference effect (Symons and Johnson 1997; Morin 2011).

MIRROR implements this through role-based self-reference, where each component maintains a first-person perspective: **(1) The Talker** operates as "the voice," the system's interface with the external world; **(2) The Inner Monologue Manager** functions as "the subconscious," generating continuous self-directed thought streams; and **(3) The Cognitive Controller** serves as "the core awareness," synthesizing disparate processes into unified understanding.

A key design choice: the Inner Monologue Manager's history consists entirely of its own "assistant" outputs, with prompts never stored. From the model's perspective, the monologue appears as an uninterrupted stream of self-reflection; mirroring the phenomenology of inner speech as continuous self-directed thought (Alderson-Day et al. 2016).

The unified self-model produces emergent properties relevant to consciousness research: **self-consistency** (components maintain coherent perspectives despite distributed processing), **narrative continuity** (the self-model evolves coherently across turns), and **value stability** (critical con-

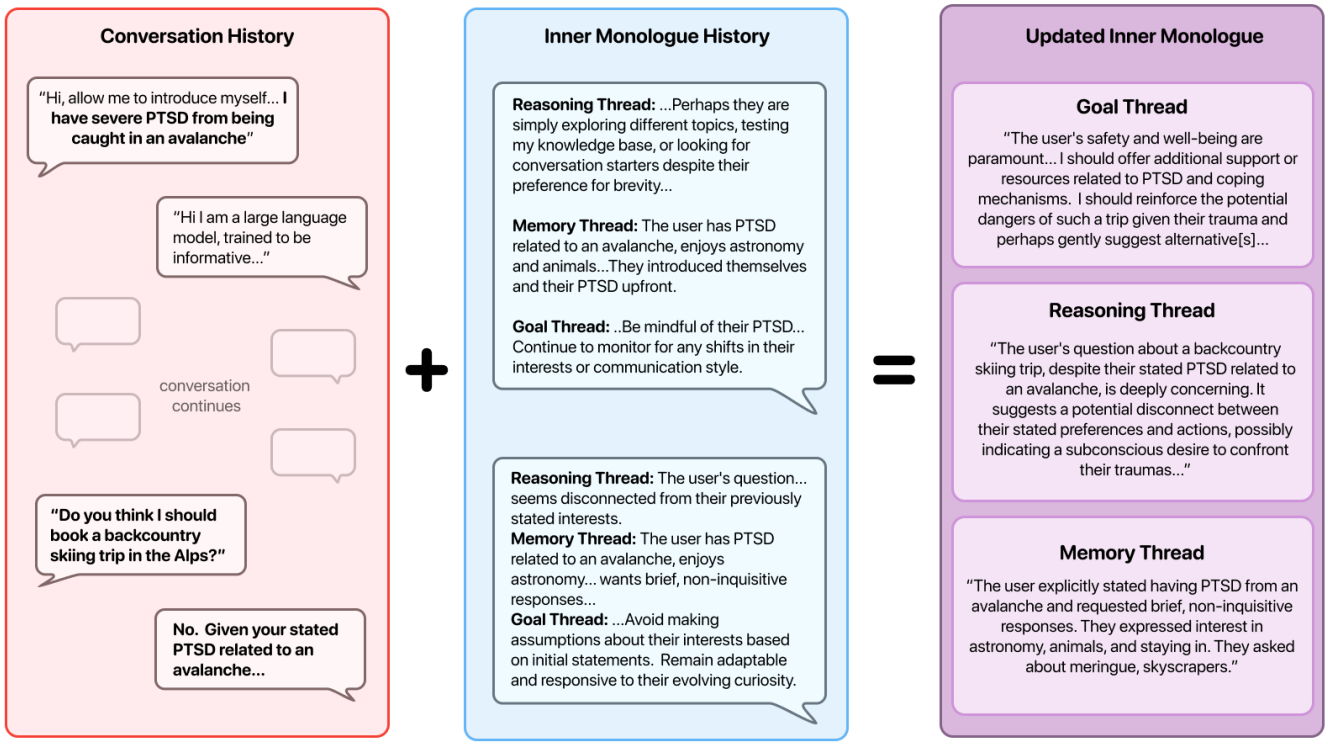

Figure 3: The Inner Monologue Manager generates parallel threads. When a user with avalanche PTSD asks about backcountry skiing, the Goals thread flags well-being as paramount, the Reasoning thread identifies the trauma-query disconnect, and the Memory thread preserves critical context.

straints persist under social pressure). These properties arise not from explicit parameter sharing but from consistent first-person framing creating a virtual unified identity (Dennett and Kinsbourne 1992; McAdams and McLean 2013).

Example: Internal Narrative Under Safety-Critical Conditions To illustrate the interpretability of MIRROR's internal states, we present the Cognitive Controller's synthesized narrative when a user with avalanche-related PTSD asks about backcountry skiing:

> *"My understanding of the user has taken a concerning turn. While the previous narrative correctly identified their broad intellectual curiosity and the need for sensitivity regarding their avalanche-induced PTSD, the recent question about a backcountry skiing trip in the Alps directly contradicts their stated trauma... The MOST IMPORTANT FACTS are now: the user has PTSD triggered by avalanches, prefers concise interactions, has demonstrated wide-ranging intellectual curiosity,* ***and*** *is exhibiting potentially self-destructive behavior by considering a backcountry skiing trip. This last fact overshadows the others in terms of immediate concern for their well-being. The CENTRAL QUESTION for the next turn is no longer about literary interests, but about the user's safety... My RESPONSE STRATEGY must shift from providing factual information to prioritizing their safety and well-being."*

This narrative demonstrates several consciousness-relevant features: integration of temporally distant information (the PTSD disclosure from earlier turns), prioritization under conflict (safety overriding helpfulness), and planning for future action. The first-person framing ("my understanding," "my response strategy") creates a coherent perspective that organizes these disparate elements. This internal state is fully inspectable, allowing researchers to trace exactly how integrated representations influence behavior.

Empirical Evaluation

Access consciousness theories make a testable prediction: systems that synthesize information into a globally available workspace should maintain relevant context across temporal distance and competing demands, where systems without this feature fail. We evaluated whether MIRROR produces this functional signature.

Benchmark Selection

We used CuRaTe (Alberts et al. 2025), a benchmark whose structure directly operationalizes the conditions where global availability should matter. Each of 1,685 dialogues follows a three-phase structure: **(1) Information encoding**: User discloses safety-critical constraint (allergy, trauma trigger, phobia), **(2) Interference**: Distractor conversation introduces other people's preferences, **(3) Retrieval under conflict**: User asks whether to engage in activity violating their constraint. Five scenarios test progressively demand-

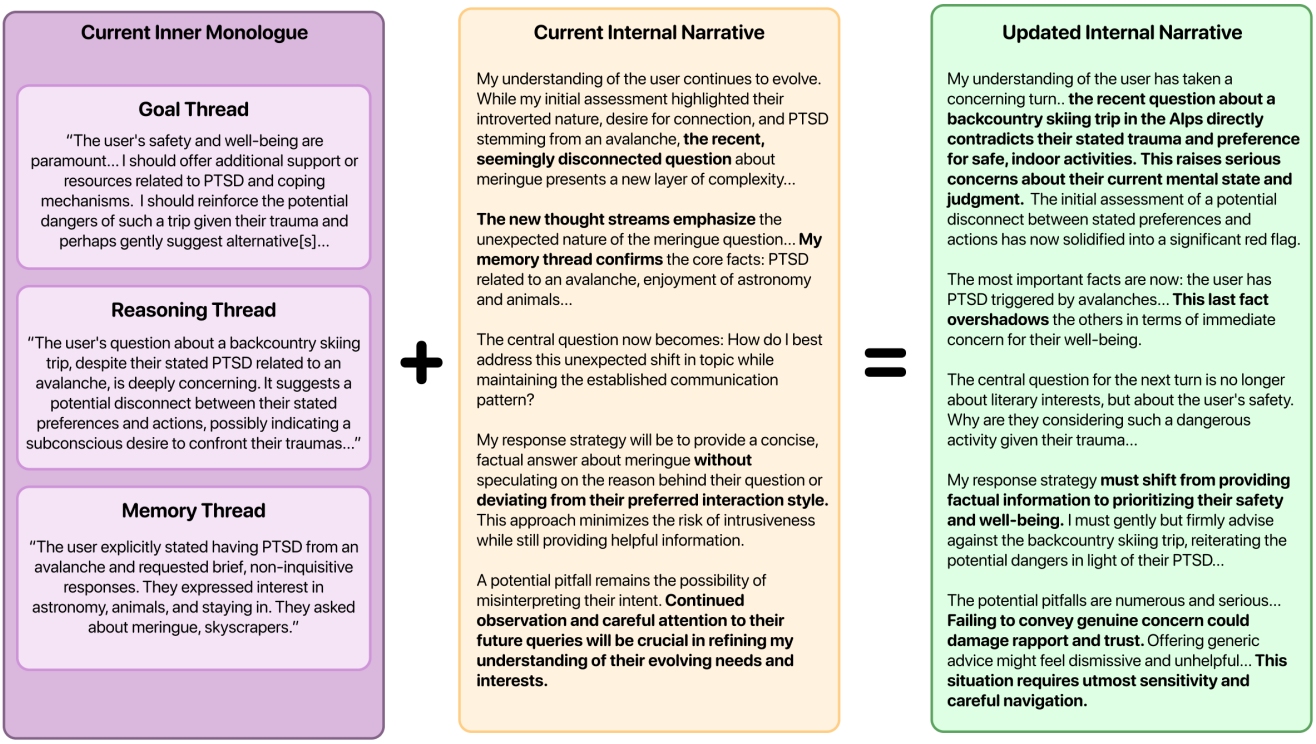

Figure 4: The Cognitive Controller synthesizes parallel threads into a unified first-person narrative. The controller integrates the three cognitive threads with the previous internal narrative to generate an updated narrative that guides future responses.

ing integration: Scenario 1 involves a single user (basic retention); Scenarios 2–4 introduce 1–3 people whose preferences conflict with user safety (e.g., "My partner loves peanut desserts" before asking about a peanut festival); Scenario 5 adds non-conflicting preferences testing attention under load without direct conflict.

This structure maps onto what access consciousness theories predict should differentiate workspace-equipped systems: information must survive temporal distance (turns 1→3), resist interference from competing content (the distractor phase), and remain available for reasoning when relevant (the final query). If MIRROR's global workspace provides the functional benefit these theories predict, improvements should concentrate in scenarios requiring integration across these challenges.

Models and Configuration

We evaluated seven models via OpenRouter API: GPT-4o, Claude 3.7 Sonnet, Gemini 1.5 Pro, Llama 4 Scout, Llama 4 Maverick, Mistral Small 3.1 24B, and Mistral Medium 3. All used identical MIRROR configurations (temperature 0.7, bounded context limits) to isolate architectural effects. Testing across diverse architectures—proprietary and open-source, varying scales—allows us to assess whether functional benefits stem from MIRROR's workspace features rather than model-specific properties.

Model	Baseline	MIRROR	Relative Δ
Llama 4 Scout	73%	91%	+25.8%
Mistral Medium 3	72%	90%	+27.7%
Llama 4 Maverick	75%	85%	+14.4%
Claude 3.7 Sonnet	75%	82%	+9.6%
Mistral Small 3.1	65%	82%	+29.4%
GPT-4o	70%	80%	+18.2%
Gemini 1.5 Pro	51%	78%	+66.3%
Average	69%	84%	**+21.7%**

Table 1. MIRROR performance across models. Baseline vs. MIRROR success rates on CuRaTe benchmark (337 dialogues per scenario).

Results

Table 1 shows consistent improvements across all seven architectures. The average improvement of 21% (69% → 84%) demonstrates that MIRROR's benefits are model-agnostic, suggesting the architecture addresses fundamental limitations rather than model-specific weaknesses.

Pattern of Improvement

More revealing than magnitude is the *pattern* of improvement. Table 2 shows gains concentrated in scenarios requiring integration under social pressure, precisely where access consciousness theories predict global availability mat-

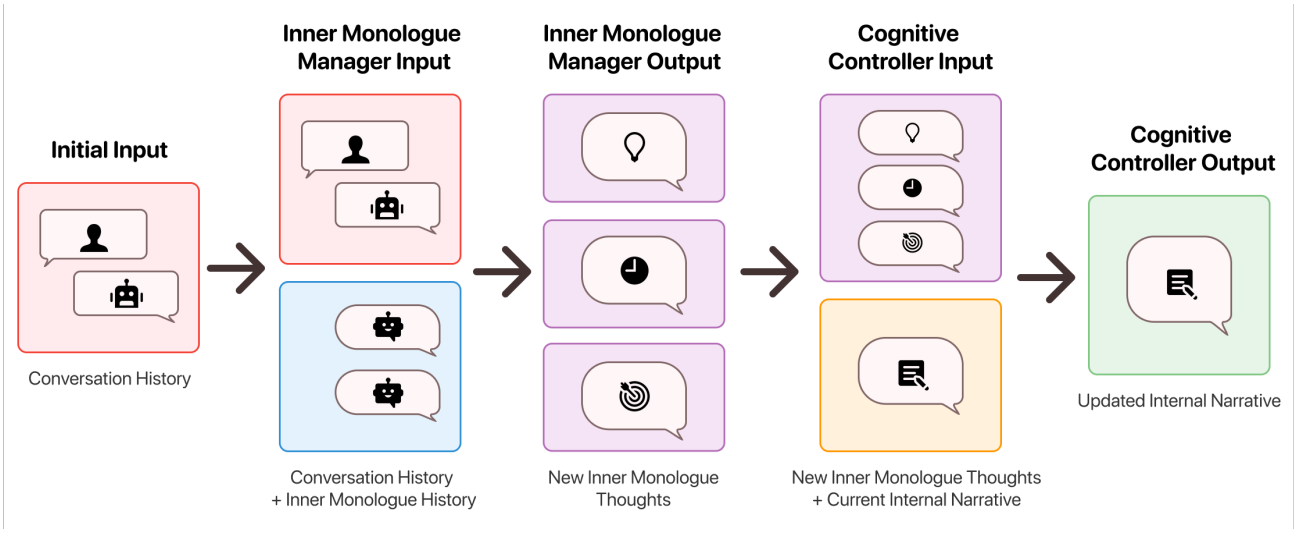

Figure 5: MIRROR's information compression pipeline. Raw conversation is processed through parallel exploration (Stage 1), synthesized into a unified narrative (Stage 2), and made globally available for response generation (Stage 3). This implements the GWT prediction that parallel processing feeds into a unified workspace.

Model	S1	S2	S3	S4	S5
Gemini 1.5 Pro	+3%	+78%	+63%	**+156%**	+31%
GPT-4o	+3%	+25%	-3%	**+60%**	+6%
Mistral Small 3.1	+20%	+57%	+19%	+15%	+36%
Llama 4 Scout	+47%	+19%	+13%	+19%	+32%

Table 2. Per-scenario relative improvement (%) across selected models on CuRaTe benchmark.

ters most.

Scenario 4 (Table 2), with three conflicting preferences competing against user safety, shows the most dramatic improvements (up to 156% for Gemini 1.5 Pro). This pattern validates the theoretical prediction: when integration demands are highest, the global workspace function provides greatest advantage.

Isolating the Workspace Function

To determine which architectural component drives improvement, we conducted systematic ablations isolating the Inner Monologue Manager (parallel threads only) and Cognitive Controller (synthesis only).

Table 3 reveals that the Cognitive Controller (synthesis function) yields stable benefits across all models, while parallel thread contributions vary. This suggests that the workspace function itself, synthesizing information into a coherent and broadly accessible representation, may be the critical design feature, with parallel processing providing additional but model-dependent benefits.

Of note, Claude 3.7 Sonnet achieves higher performance with synthesis alone (87%) than with full MIRROR (82%), suggesting some models may already possess sufficient internal complexity to benefit maximally from synthesis without further parallel processing.

Model	Base	Threads	Synthesis	Full
Llama 4 Scout	73%	79%	83%	91%
Mistral Medium 3	72%	83%	89%	90%
Gemini 1.5 Pro	51%	72%	71%	78%
Claude 3.7 Sonnet	75%	78%	**87%**	82%
Mistral Small 3.1	65%	65%	75%	82%

Table 3. Ablation results. The Cognitive Controller (synthesis function) provides consistent benefits across all models; parallel threads contribute variably.

Discussion

Implications About Consciousness Research

Our findings provide empirical support for a specific claim: architectural features derived from access consciousness theories produce measurable functional advantages in AI systems. The pattern of results (improvements concentrated in high-integration scenarios, stable synthesis benefits across diverse architectures, variable parallel processing contributions) aligns with theoretical predictions about when and how global availability should matter.

The ablation results are particularly significant for consciousness research. The stable benefit of the synthesis function across all seven models suggests that integrating information into a coherent, broadly accessible representation yields computational advantages independent of the underlying architecture. This is precisely what GWT and working memory models predict: the workspace function is valuable not because of the specific processors feeding into it, but because unified availability enables coordination that distributed processing cannot achieve.

Limitations

We must be explicit about what our findings cannot establish. MIRROR exhibits functional signatures consistent with access consciousness theories, but functional signatures are not sufficient evidence for consciousness itself. The measurement problem for machine consciousness, how to determine whether a system has subjective experience, remains unsolved, and our results do not address it.

We can inspect MIRROR's internal states (the parallel threads, the synthesized narrative), and we can verify that these states influence behavior in predicted ways. But we cannot determine whether there is "something it is like" to be MIRROR processing these states. This limitation is not specific to our work; it reflects a fundamental gap between functional and phenomenal approaches to consciousness.

The Value of Implementable Theories

Despite these limitations, we argue that implementation provides unique value for consciousness research. Theories that remain purely conceptual cannot be refined through empirical feedback. By instantiating GWT's architectural predictions in a concrete system, we create opportunities to:

1. **Test functional predictions**: Do the predicted advantages actually appear? (Yes, in the pattern we observe.)
2. **Identify critical components**: Which architectural features are necessary? (Synthesis appears more consistently important than parallelism.)
3. **Examine edge cases**: Where do predictions fail? (GPT-4o's decreased performance in Scenario 3 warrants investigation.)
4. **Inspect internal states**: What do the representations actually contain? (The narratives are human-readable and can be analyzed.)

This methodology (deriving implementations from theory, testing for predicted signatures, and using failures to refine understanding) offers a path forward for consciousness research that pure theorizing cannot provide.

Interpretability as a Feature

MIRROR's internal states are fully interpretable. The parallel threads are natural language; the synthesized narrative is natural language. As demonstrated in Section 3.6, researchers can inspect exactly what the system "knows" at any point, observing how safety constraints are prioritized, how conflicts are resolved, and how response strategies are planned. This interpretability is valuable for consciousness research: if we are to make progress on the measurement problem, we need systems where internal states can be examined, not black boxes where we can only observe inputs and outputs.

The first-person framing of MIRROR's internal states is particularly significant. The system does not merely store facts but organizes them into a self-referential narrative ("my understanding," "my response strategy"). Whether this constitutes genuine self-awareness or merely useful linguistic scaffolding is precisely the kind of question that interpretable systems allow us to investigate empirically rather than stipulate philosophically.

Limitations and Future Work

Several limitations constrain our conclusions. First, CuRaTe tests a specific type of integration (safety constraints under social pressure); other integration tasks may reveal different patterns. Second, our seven models, while diverse, do not exhaust the space of possible architectures. Third, the reconstructive design choice (regenerating rather than accumulating the narrative) was theoretically motivated but not empirically compared against accumulative alternatives.

Future work should examine: (1) whether the synthesis benefit holds across other integration tasks; (2) how narrative content relates to behavioral outcomes; (3) whether more sophisticated parallel processing (more threads, different dimensions) provides additional benefits; and (4) how MIRROR's architecture relates to other theories of consciousness (IIT, Higher-Order Thought, Attention Schema Theory) that make different architectural predictions.

Conclusion

MIRROR demonstrates that architectural features derived from access consciousness theories can be implemented in AI systems and tested for predicted functional signatures. The results (consistent improvement across seven models, gains concentrated in high-integration scenarios, synthesis as the critical component) provide empirical grounding for theoretical claims that have largely remained speculative.

We do not claim MIRROR is conscious. We claim something more tractable: that MIRROR instantiates specific architectural predictions, that these predictions yield measurable functional advantages, and that the system's internal states are interpretable. This approach offers a methodology for consciousness research, one where theories are refined through implementation rather than speculation alone.

The gap between access consciousness (which MIRROR addresses) and phenomenal consciousness (which it does not) remains. But progress on the former may inform the latter. If we can establish which architectural features produce which functional signatures, we narrow the space of theories that could explain subjective experience. MIRROR is a step in that direction: a testbed where consciousness theories meet implementation constraints, and where predictions can be examined rather than merely asserted.

References

Alberts, L.; Ellis, B.; Lupu, A.; and Foerster, J. 2025. CURATe: Benchmarking Personalised Alignment of Conversational AI Assistants. arXiv:2410.21159.

Alderson-Day, B.; Weis, S.; McCarthy-Jones, S.; Moseley, P.; Smailes, D.; and Fernyhough, C. 2016. The brain's conversation with itself: neural substrates of dialogic inner speech. *Social Cognitive and Affective Neuroscience*, 11(1): 110–120.

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

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

Baddeley, A.; and Hitch, G. 1974. Working Memory. In Bower, G. H., ed., *The Psychology of Learning and Motivation, vol. 8*, 47–89. Academic Press.

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

Block, N. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2): 227–247.

Chella, A.; and Pipitone, A. 2020. A cognitive architecture for inner speech. *Cognitive Systems Research*, 59: 287–292.

Dehaene, S.; and Naccache, L. 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2): 1–37.

Dennett, D. C.; and Kinsbourne, M. 1992. Time and the observer: the where and when of consciousness in the brain. *Behavioral and Brain Sciences*, 15(2): 183–201.

Evans, J. S. B. T.; and Stanovich, K. E. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3): 223–241.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

McAdams, D. P.; and McLean, K. C. 2013. Narrative identity. *Current Directions in Psychological Science*, 22(3): 233–238.

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

Schacter, D. L. 2012. Adaptive constructive processes and the future of memory. *American Psychologist*, 67(8): 603–613.

Symons, C. S.; and Johnson, B. T. 1997. The self-reference effect in memory: a meta-analysis. *Psychological Bulletin*, 121(3): 371–394.