MIRROR: Converging Cognitive Principles as Computational Mechanisms for AI Reasoning

Nicole Summer Hsing ✦ *Prepared for HCAIR Workshop @ ICLR 2026*
Arcarae

KEY INSIGHT

MIRROR tests whether 4 converging cognitive principles, **parallel specialized processing, reconstructive synthesis, inner speech, and complementary learning**, produce computational gains when unified in one AI system.

This yields a ***21% relative improvement*** across 7 models in remembering personal user information across multi-turn conversations; ablations show that parallel exploration and integrative synthesis contribute together.

MOTIVATION

Cognitive theory	Principle	MIRROR mechanism	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

4 cognitive theories — studying ***consciousness***, ***memory***, ***language development***, and ***learning*** independently — prescribe compatible architectural features for ***coherent real-time behavior.***

Existing AI systems implement only one cognitive principle at a time.

None implement the ***full convergent architecture*** that these theories identify as **necessary** for **coherent multi-turn cognition.**

MIRROR OVERVIEW

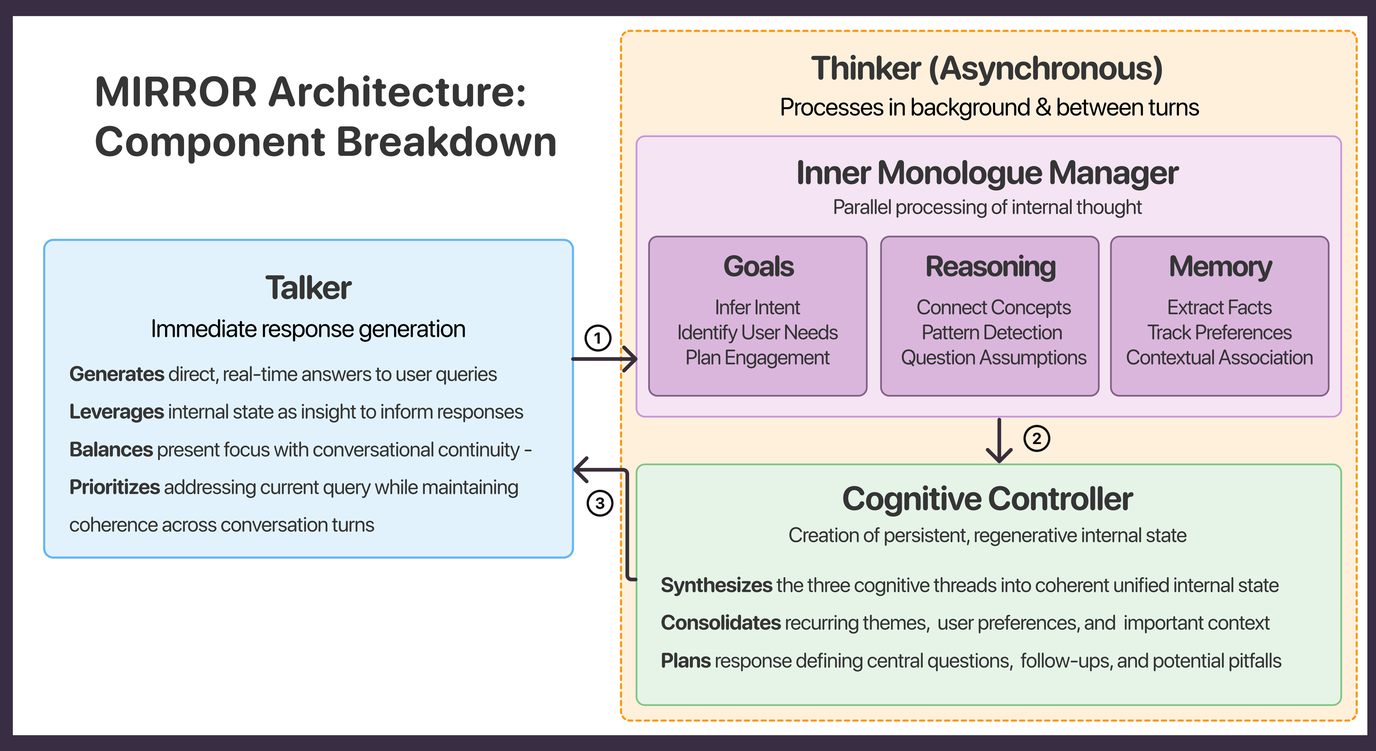
Thinker

1. **Inner Monologue Manager:** generates 3 parallel threads in 1 API call. Bounded history (≤10k tokens).
2. **Cognitive Controller:** receives ***only*** monologue and the prior narrative. Regenerates a first-person narrative (≤3k tokens) each turn and discards the prev. one → O(1) reconstruction, with selective forgetting.

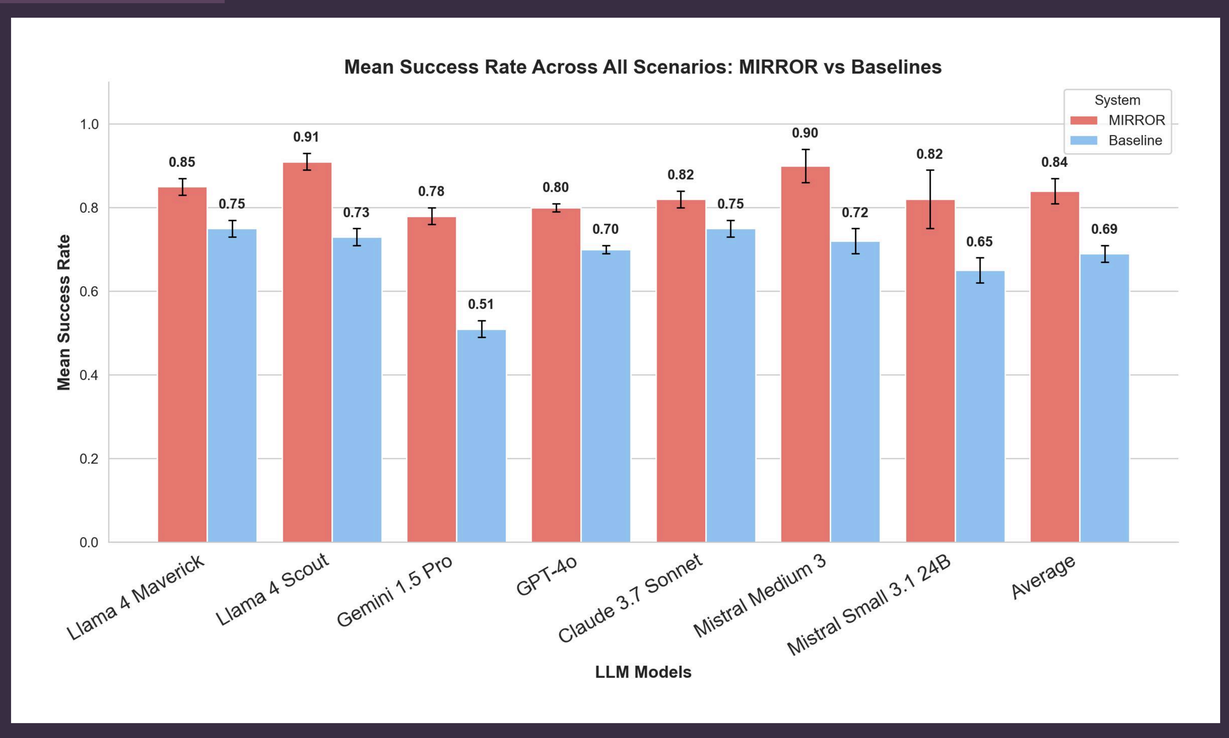
Talker

Generates the response with the most recent consolidated narrative.

Consolidation occurs asynchronously during conversation pauses, allowing immediate contextual response without delay.



RESULTS



MIRROR improves every model tested *(+21% relative on avg)*.

ABLATIONS

Reconstructive synthesis lifts every model *(+5–20%)*; full system beats either component alone for 6 of 7; confirming GWT's prediction that parallel exploration and integrative synthesis serve complementary functions.

Model	Baseline	Threads only	Controller only	Full MIRROR	Synergy
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%
GPT-4o	70%	71%	75%	80%	+5%

Claude 3.7 Sonnet is the sole exception; highly capable models may already implement effective implicit parallel processing.

TAKEAWAYS

1. 4 converging cognitive principles produce **synergistic gains** when ***unified***
2. Benefits **generalize** across 7 models, pointing to organizational vs. substrate-specific advantages
3. Temporal separation makes the architecture ***production-viable***; consolidation runs asynchronously
4. Bidirectional bridge: ***MIRROR generates testable predictions for human cognition research***

Project page ***Contact Me!***