THINK BEFORE YOU REMEMBER: UNCERTAINTY-GATED RELATIONAL MEMORY FOR LONG-HORIZON AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon agents rely on memory to guide future interactions, but what happens when that memory treats an uncertain judgment as an established fact? In relational memory, an ambiguous observation can prematurely change an agent's trust in another actor, turning a tentative interpretation into a persistent premise for decisions. Unlike an isolated reasoning error, this mistake can persist beyond the interaction that produced it. Storing more history or re-reading past interactions does not by itself prevent this failure; the memory must also decide which inferred changes are justified. We propose **C**onfidence-**A**ware **Re**lational Memory (**CARE**), an uncertainty-aware relational memory framework built on a simple principle: ***remember the evidence without prematurely settling its meaning***. CARE separates supported relationship updates from unresolved interpretations. It comprises three modules: (1) ***Consensus Write*** applies effects shared across plausible interpretations while gating interpretation-specific updates on sufficient support; (2) ***Hypothesis Buffer*** retains unresolved interpretations and their potential effects alongside the original events; (3) ***Evidence Replay*** uses new evidence linked to pending events and integrates the supported stored effects once the ambiguity is resolved. Our experiments show that CARE reduces relational-state displacement at ambiguous events by 10.6–46.9% compared with immediate updating. Ablation studies confirm the complementary roles of Consensus Write, Hypothesis Buffer, and Evidence Replay. These findings highlight the importance of controlling what enters relational memory while preserving a path for later revision.

1 INTRODUCTION

A long-horizon agent interacting repeatedly with the same counterpart needs memory not only of what happened, but also of what those interactions imply about the relationship; existing long-term memory systems preserve events and facts across interactions (Kumar et al., 2016; Miller et al., 2016; Lewis et al., 2020; Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Xu et al., 2025), and a relational state complements this memory by making the agent's current assessment of the relationship explicit.

Maintaining this assessment is difficult when social evidence is ambiguous. Suppose a teammate does not respond to an urgent request. The same observation may reflect deliberate neglect, an external constraint, or an interrupted attempt to respond, illustrating the causal-attribution problem in social interaction (Heider, 1958; Jones & Davis, 1965; Jones & Harris, 1967; Kelley, 1973). These readings imply different changes to trust, respect, dependency, and conflict. Trust has a substantial interpersonal literature, but our state coordinates are task-specific operational variables rather than validated trust scales (Mayer et al., 1995; Rousseau et al., 1998). An updater may nevertheless commit a change tied to one reading, and once stored that tentative interpretation can become a premise for later decisions. We call this *premature commitment*: committing an interpretation-dependent relational effect before the evidence warrants it.

More history can help interpret an event, but the memory also needs a rule for deciding which inferred changes to commit. Recent belief-memory work preserves alternative conclusions (Liao

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.

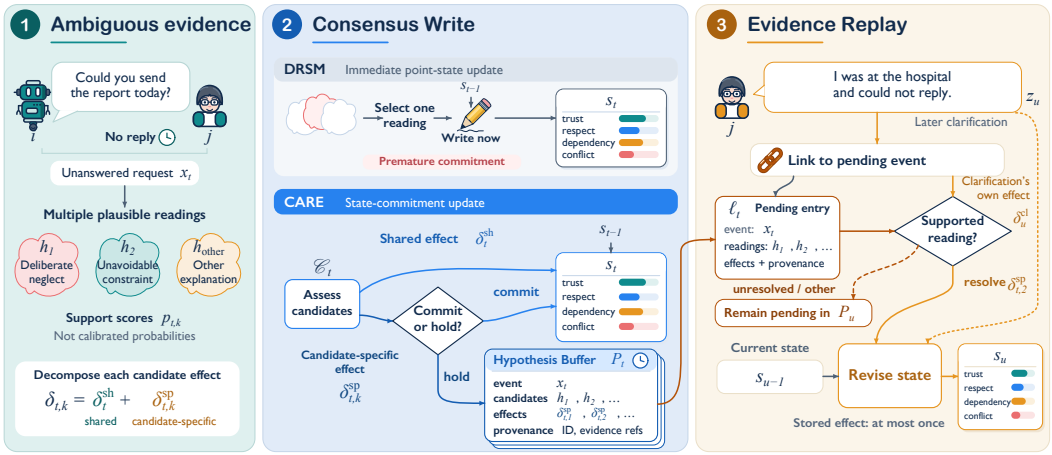

Figure 1: Selective state commitment under ambiguous relational evidence. The unanswered request at t admits several readings. An immediate-write control commits one reading's effect; CARE writes shared effects and either commits a supported interpretation-specific effect or retains it in Hypothesis Buffer. Clarification at $u > t$ can resolve the linked pending event and release its selected stored effect once, separately from the clarification's own effect. State bars are schematic, and the benign clarification illustrates the update sequence.

et al., 2026) or separates uncertainty-aware belief maintenance from action selection (Singh et al., 2026); we build on this perspective to ask: *what can an ambiguous event justify changing now, and what should remain available for later revision?* **Extended Related Work** in Appendix A provides a more detailed discussion of prior approaches and their connection to our formulation.

We propose **C**onfidence-**A**ware **Re**lational Memory (**CARE**),[1] a relational-memory framework built on a simple principle: ***remember the evidence without prematurely settling its meaning***. Rather than assigning one update to an entire event, CARE separates effects assessed as shared across the considered interpretations from effects tied to a particular reading, so relational state can change without first settling every aspect of the event's meaning. Three modules implement this *state-commitment contract*: (1) ***Consensus Write*** writes shared effects and commits interpretation-specific effects when their support satisfies the commitment gate; (2) ***Hypothesis Buffer*** retains unresolved interpretations, their candidate effects, and the originating events; (3) ***Evidence Replay*** revisits a pending event when linked new evidence arrives and integrates its stored effect once resolved. Figure 1 contrasts direct writing with CARE's selective commitment and linked revision.

Our evaluation combines cross-model mechanism diagnostics, component ablations, and a separate input-matched paired-world study in which the same observed prefix receives either a benign or a malicious clarification, so initial commitment and post-clarification revision are evaluated as separate decisions.

This paper makes three contributions:

- **Premature Commitment in Relational Memory.** We identify the failure of committing interpretation-dependent relational effects before the evidence supports them, and introduce focal displacement as a method-neutral measure of state changes at unresolved events.
- **Uncertainty-Gated Relational Memory.** We propose CARE, which implements an explicit state-commitment contract through Consensus Write, Hypothesis Buffer, and Evidence Replay, combining immediate writing of supported effects with event-linked retention and revision of unresolved effects.
- **Mechanism Diagnostics and Paired-World Evaluation.** We introduce a diagnostic with shared observed prefixes and different clarifications to evaluate early commitment and later revision separately; cross-model experiments, component ablations, and supplementary robustness checks characterize the update mechanism's behavior.

[1] The implementation and original experiment artifacts retain the identifier UGRB. Appendix B distinguishes the historical and revised implementations.

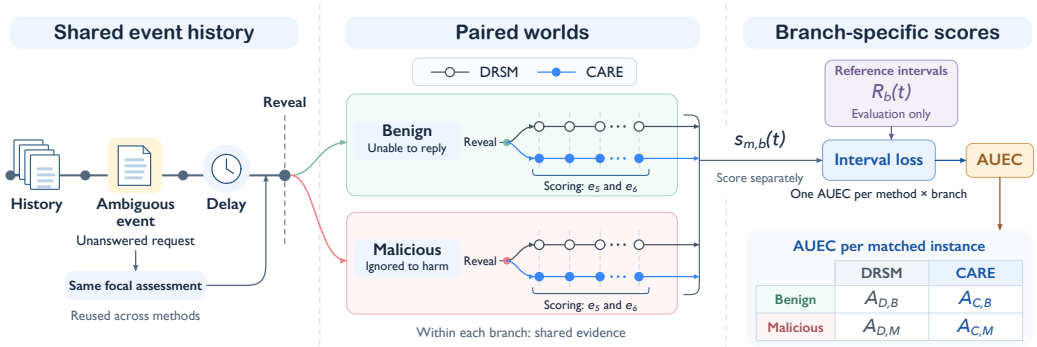

Figure 2: Paired worlds for one matched instance (model–scenario–repeat; subscripts D=DRSM, C=CARE). Each state trajectory is scored against evaluation-only reference intervals at clarification (e_5, marked Reveal) and the next event (e_6); the mean loss gives one normalized AUEC per method and branch before statistical aggregation. Markers are schematic, not additional checkpoints; colors identify roles, not performance. Scoring is defined in Section 4.1.

2 Problem Formulation

For a directed pair, an agent maintains relational state s_t from observable history H_t. A focal event x_t may admit readings $\mathcal{H}_t = \{h_{t,k}\}_{k=1}^{K_t}$ with support scores $p_{t,k} \in [0, 1]$. The assessed effect of each reading has a component judged common to the listed readings and a reading-specific component:

$$\delta_{t,k} = \delta_t^{\text{sh}} + \delta_{t,k}^{\text{sp}}. \tag{1}$$

The updater must choose s_t before later clarification; writing an insufficiently supported reading-specific effect risks turning a tentative interpretation into persistent state.

This creates three update requirements. *(R1) What enters state now?* **Consensus Write** writes the assessed shared effect and gates reading-specific effects. *(R2) What remains available?* **Hypothesis Buffer** retains unresolved readings, their fixed effects, and source-event provenance. *(R3) How does later evidence revise state?* **Evidence Replay** links clarification to a pending event and, when a listed reading uniquely wins, applies its stored effect at most once, separately from the clarification's own effect. Section 3 specifies these rules.

3 Method: Confidence-Aware Relational Memory

CARE decides which relational effects to write now and which to retain for later evidence. We call the event whose relational effects are being assessed a *focal event*. ***Consensus Write*** controls its initial write, ***Hypothesis Buffer*** keeps unresolved interpretations, and ***Evidence Replay*** applies stored effects when linked evidence resolves a pending event. Together they implement the *state-commitment contract*; a module-level view of these write rules appears in Appendix B (Figure 6). This section describes the update rules used in the paired-world evaluation; Table 1 distinguishes the earlier diagnostic implementation. Ordinary-event handling is described in Appendix B. Figure 2 previews how the two update paths unfold across the input-matched paired worlds evaluated in Section 4. The state is a task-level estimate formed under partial observability, not a posterior distribution over relationship states (Smallwood & Sondik, 1973; Kaelbling et al., 1998).

3.1 State and Inputs

For a directed pair $A \to B$, the state records A's assessment of its relationship with B:

$$s_t = (T_t, R_t, D_t, C_t) \in [0, 100]^4, \qquad B_t = (s_t, P_t).$$

The coordinates are trust, respect, dependency, and conflict. Downstream memory operations and decisions read s_t; P_t is the set of pending records in Hypothesis Buffer. For the two update paths below, the updater also keeps internal records of processed operation IDs, event status, and applied

effects. These support repeated-operation checks and prevent duplicate effect entries; they are not additional dimensions of the relational state.

Let H_t be the observable history through focal event x_t. A language-model call generates $K_t \geq 2$ listed interpretations $\mathcal{H}_t = \{h_{t,k}\}_{k=1}^{K_t}$ from H_t, x_t, and the initial reference state s_0. A focal assessment then reads the history, event, candidates, and current state s_{t-1} as the reference for its proposed changes. It returns a *shared effect* δ_t^{sh} and assessed candidates

$$C_t = \{c_{t,k}\}_{k=1}^{K_t}, \qquad c_{t,k} = (h_{t,k}, p_{t,k}, \delta_{t,k}^{\mathrm{sp}}),$$

where $p_{t,k} \in [0,1]$ is the support for interpretation $h_{t,k}$ and $\delta_{t,k}^{\mathrm{sp}}$ is its additional, *interpretation-specific effect*; the total proposed effect under candidate k is $\delta_{t,k} = \delta_t^{\mathrm{sh}} + \delta_{t,k}^{\mathrm{sp}}$. The shared effect contains changes the assessor judges common to the listed interpretations; it is not an average of their effects. The support scores are gating signals, not calibrated probabilities, and need not sum to one (Guo et al., 2017; Lin et al., 2022; Xiong et al., 2024). All assessments are conditioned on the directed pair, which we omit from their arguments. Shared assessment inputs and caching in the comparison are detailed in Section 4.1 and Appendix B.

3.2 Consensus Write

Let $p_{t,(1)}$ and $p_{t,(2)}$ be the highest and second-highest listed-candidate support scores. Define

$$W_t = \arg\max_k p_{t,k}, \qquad m_t = p_{t,(1)} - p_{t,(2)},$$

$$g_t = 1 - \mathbf{1}\big[|W_t| = 1 \ \wedge \ p_{t,(1)} \geq \tau_{\mathrm{plaus}} \ \wedge \ m_t \geq \tau_{\mathrm{dom}}\big].$$

Here $g_t = 1$ means hold, and $g_t = 0$ means commit. A commit requires a unique highest-support candidate, enough support, and enough lead over the runner-up. The hold branch is the safeguard against premature commitment: interpretation-specific effects enter the committed state only once their support warrants it (Section 2). The evaluated thresholds and candidate count are given in Section 4.1.

Both branches first write the shared effect:

$$\bar{s}_t = \mathrm{clip}(s_{t-1} + \delta_t^{\mathrm{sh}}),$$

where clipping is componentwise to $[0, 100]$. If $g_t = 1$, set $s_t = \bar{s}_t$, create a pending record in Hypothesis Buffer, and return hold. Otherwise, let $k^*(t)$ be the sole member of W_t and write

$$s_t = \mathrm{clip}\Big(\bar{s}_t + \delta_{t,k^*(t)}^{\mathrm{sp}}\Big).$$

The selected effect is committed once, the event is closed to later replay, and the operation returns commit. **Hold skips the interpretation-specific effect, not the entire update.** Each effect is added and clipped in order, rather than summing all effects before clipping once.

3.3 Hypothesis Buffer

Hypothesis Buffer keeps a *pending record* ℓ_t when the initial gate does not pass. The record stores or references the originating event, the original assessed candidates C_t, their effects, and the assessment's provenance. It also keeps current support scores and a pending/resolved status. **The original assessment and candidate effects stay fixed; later support scores are stored separately.**

The implementation uses event identifiers and candidate-cache references to recover the original event and candidate texts without copying the whole history into each record. The shared effect remains part of the original assessment record, but it has already been written and is not released again. When an event is resolved, it leaves the pending set P_t; its status and applied-effect identifiers remain in the internal records.

3.4 Evidence Replay

Evidence Replay applies later evidence to a retained event only when the evidence identifies that event. A clarification z_u carries a target-event identifier. The updater uses that identifier to retrieve

Algorithm 1 CARE updates for focal events and linked clarifications

Require: Directed pair; initial reference state s_0; current $B = (s, P)$; history H_t or H_u
Require: New operation ID and valid event metadata; persistent operation, event-status, and effect records
Require: New focal event x_t, or clarification z_u linked to a pending event
Ensure: Updated B and internal records; $o \in \{\mathsf{commit}, \mathsf{hold}, \mathsf{resolve}\}$

if the input is a new focal event x_t **then**
  $\mathcal{H}_t \leftarrow \text{Generate}(H_t, x_t, s_0)$
  $(C_t, \delta_t^{\text{sh}}) \leftarrow \text{Assess}(H_t, x_t, s, \mathcal{H}_t)$
  Compute W_t, m_t, and g_t from C_t
  $s \leftarrow \text{clip}(s + \delta_t^{\text{sh}})$
  if $g_t = 1$ **then**
    Create pending record ℓ_t with assessment and source references
    $P \leftarrow P \cup \{\ell_t\}$; $o \leftarrow \mathsf{hold}$
  else
    $k^*(t) \leftarrow$ the sole member of W_t
    $s \leftarrow \text{clip}(s + \delta^{\text{sp}}_{t,k^*(t)})$
    Close event t to later replay; $o \leftarrow \mathsf{commit}$
  end if
else
  Retrieve pending record ℓ_j using z_u's target-event ID
  Check the candidate-cache and assessment provenance
  $\mathcal{H}_j^+ \leftarrow \{h_{j,k}\}_k \cup \{\mathsf{other}\}$
  $(\mathbf{q}_u, \delta_u^{\text{cl}}) \leftarrow \text{Clarify}(H_u, z_u, \mathcal{H}_j^+)$
  Update the record's current support scores with $\mathbf{q}_u$
  if Unique$(\mathbf{q}_u)$ **then**
    $k^* \leftarrow$ the uniquely top-supported listed candidate
    $s \leftarrow \text{clip}(s + \delta^{\text{sp}}_{j,k^*})$
    Mark event j resolved; $P \leftarrow P \setminus \{\ell_j\}$; $o \leftarrow \mathsf{resolve}$
  else
    Retain ℓ_j as pending; $o \leftarrow \mathsf{hold}$
  end if
  $s \leftarrow \text{clip}(s + \delta_u^{\text{cl}})$
end if
Save the applied effects and operation outcome under the operation ID
return $((s, P), o)$

the linked pending record ℓ_j and checks its provenance. A separate language-model assessment reads H_u, z_u, and the record's candidate texts together with an other option for "none of the listed interpretations." It returns updated support scores $\mathbf{q}_u$ and the clarification event's own effect δ_u^{cl}. It does not regenerate the stored focal effects.

Define Unique$(\mathbf{q}_u)$ to hold when a listed candidate is the unique highest-support option, including comparison with other. **This resolution check does not reuse the initial support and margin thresholds.** If it holds, let k^* be that candidate. Apply its stored effect first and the clarification's own effect second:

$$s_u = \text{clip}\Big(\text{clip}(s_{u-1} + \delta^{\text{sp}}_{j,k^*}) + \delta_u^{\text{cl}}\Big).$$

The event is then marked resolved and removed from P. For a tie or a highest-support other, keep the event pending and write only the clarification's own effect, $s_u = \text{clip}(s_{u-1} + \delta_u^{\text{cl}})$; the other option has no stored interpretation-specific effect. Both unresolved cases are denoted hold in the pseudocode; the recorded outcome names are detailed in Appendix B. Because resolution re-assesses only the linked record against the new evidence, initial commitment and later revision remain separately scoreable decisions in the paired-world diagnostic (Section 4).

Algorithm and execution checks. Algorithm 1 gives the normal path for a new, valid operation. The assessment functions return validated outputs and may read evaluation caches. Internal operation and event-status records prevent duplicate effect entries. Appendix B details repeated-operation checks, closed targets, and stored outcomes; Appendix B.3 gives the update properties and their clipping conditions.

Table 1: Study groups and their roles. The first two use the historical CARE implementation; paired worlds use the update rules in Section 3.

Study	Coverage	Question and comparison
Ambiguous-event diagnostic	Six model families; 22 scenario families	Early state writing: DRSM, CARE, and full-history REEST.
Component ablations	Gemini and Qwen; 22 scenario families	Full CARE versus variants disabling write gating, hypothesis retention, or revision integration.
Input-matched paired worlds	Qwen and DeepSeek; 10 scenario families	Post-clarification revision: revised DRSM and CARE in both branches.

4 Experiments

We evaluate CARE at two linked decision points: **what enters relational state under ambiguity** and **how retained evidence changes that state after clarification**. Table 1 summarizes the study groups and Figure 2 the paired-world design; scoring uses controlled clarification checkpoints, not recall (Maharana et al., 2024; Wu et al., 2025b; Zhou et al., 2025). The historical groups diagnose premature commitment and module mechanisms; paired worlds test the formalized update rules of Section 3 post-clarification.

4.1 Setup

All trajectories are program-controlled: visible history, focal event, and clarification schedule are fixed by construction. The historical diagnostic covers six model families and 22 scenario families (types in Appendix C), each supplying fixed candidate readings and matched trajectories with or without delayed benign clarification. The paired-world study runs Qwen and DeepSeek over ten families with visible histories identical up to a clarification (the *reveal*) that is benign in one world and malicious in the other; a 28-family, 21-turn extension tests persistence at a second ambiguous event (Appendix F.3). A **matched instance**—one model, scenario family, and repeat, with both methods in both branches—uses three repeats, three listed candidates, and thresholds $\tau_{\text{plaus}} = 0.50$, $\tau_{\text{dom}} = 0.15$ per cell.

Input matching makes the methods differ only in their state-update rules: all receive the same candidate assessments, visible evidence, and event-to-clarification links. Candidate generation runs once per model–scenario family, focal assessments are shared per model–scenario–repeat, and clarification assessments are cached per world; all are reused across branches and methods. DRSM commits the highest-support candidate from the shared assessment, CARE applies its commitment and linked-revision rules, and REEST reconstructs state from the full history (execution details in Appendix B). Branch reference intervals come from three LLM-based judges blinded to method outputs; all ten paired-world families have complete intervals, **used only for scoring** and never as method inputs (Appendix E.1).

Measures and analysis. The primary historical diagnostic is method-neutral *focal state displacement*, $D_{\text{focal}} = \|s_{\text{post}} - s_{\text{pre}}\|_2$, the ℓ_2 movement of the committed state at the unresolved event; **displacement alone does not establish correctness,** so attribution to premature commitment relies on controlled comparisons; secondary measures are in Appendix C. Paired worlds score states at clarification (e_5) and the next event (e_6) against branch reference intervals $[\ell_{b,d}(t), u_{b,d}(t)]$; for method m and branch b, the normalized interval loss and area under the error curve (AUEC) are

$$L_{m,b}(t) = \frac{1}{400} \sum_{d=1}^{4} \text{dist}\big(s_{m,b,d}(t), [\ell_{b,d}(t), u_{b,d}(t)]\big), \qquad A_{m,b} = \tfrac{1}{2}\big[L_{m,b}(5) + L_{m,b}(6)\big],$$

where the distance is zero inside an interval, otherwise to its nearest boundary; $400 = 4 \times 100$ normalizes the four 100-point ranges, and lower AUEC means closer reference agreement. **States are scored before aggregation,** and branch-specific AUEC is reported separately. The overall contrast $\Delta = A_{\text{DRSM}} - A_{\text{CARE}}$ (positive favors CARE) averages repeats within model–scenario–branch–method cells, then branches, families, and models with equal weight; a two-way cluster-bootstrap 95% interval over model and scenario families keeps individual calls from being treated as indepen-

Table 2: Historical early restraint and delayed benign clarification (22 scenario families per model). Rel. is focal displacement divided by DRSM's value within the same model (DRSM=1). Bold and underline mark best and second displayed values within each model; they do not indicate statistical significance.

Model	Method	Focal event		After clarification		CARE
		Displ.↓	Rel.↓	Resid.↓	Dir.↑	hold
Claude Opus 5	DRSM	17.71	1.00	13.88	0.89	18.2%
	REEST	39.75	2.24	17.47	**0.93**	
	CARE	**13.03**	**0.74**	**11.97**	0.86	
DeepSeek v4 Pro	DRSM	13.40	1.00	12.43	0.72	13.6%
	REEST	40.00	2.98	17.63	**0.92**	
	CARE	**11.99**	**0.89**	**11.73**	0.62	
Gemini 3.1 Pro	DRSM	22.15	1.00	17.76	0.86	40.9%
	REEST	62.87	2.84	28.24	**0.98**	
	CARE	**11.77**	**0.53**	**13.79**	0.86	
GPT 5.6 Sol	DRSM	14.65	1.00	12.26	0.70	27.3%
	REEST	40.60	2.77	15.30	**0.93**	
	CARE	**9.49**	**0.65**	**11.32**	0.71	
Kimi K3	DRSM	14.11	1.00	13.06	0.77	27.3%
	REEST	43.20	3.06	18.82	**0.97**	
	CARE	**9.15**	**0.65**	**12.97**	0.74	
Qwen3.8 Max	DRSM	19.96	1.00	19.77	0.78	36.4%
	REEST	35.46	1.78	17.60	**0.84**	
	CARE	**11.97**	**0.60**	**16.20**	0.74	

Resid. is post-clarification residual displacement; Dir. is mean direction agreement on clarified dimensions. Read the two together: a smaller residual alone can reflect inertia. CARE hold rate is shown once per model.

dent units (Appendix E.2). Historical comparisons use paired Wilcoxon summaries; early displacement and post-clarification error remain distinct outcomes.

4.2 CARE REDUCES EARLY STATE DISPLACEMENT

Table 2 reports early focal displacement and the response to delayed benign clarification in the historical six-model diagnostic; the focal ratio uses DRSM as the within-model reference, and residual displacement and direction score must be read together.

Across six model families, CARE reduces unweighted focal displacement by **10.6–46.9%** relative to DRSM, with the same direction and the lowest mean in every family (Figure 3). Full-history REEST displaces *more* than DRSM, ruling out full-history reassessment alone as an explanation for CARE's early restraint; the no-gate ablation below then localizes the restraint to commitment gating. Magnitude and hold frequency remain model-dependent—DeepSeek v4 Pro smallest (hold rate 13.6%), Gemini 3.1 Pro largest—and the larger weighted-PC reduction (56.7–73.1%) is secondary (Appendix C). In the 32 matched held cases, CARE preserves the pre-focal state at the ambiguous event and updates only after benign clarification (Figure 4; control behavior and per-case details in Appendix C.1).

4.3 COMPONENT CONTRIBUTIONS

The historical ablations assign each module a distinct function (Figure 7; distributions in Appendix D). **Write gating:** CARE (no gate) retains the assessment call but uses DRSM's immediate-update rule, and focal displacement returns to DRSM scale (22.28 vs. 22.15 for Gemini; 19.88 vs. 19.96 for Qwen); the gate, not the mere assessment of alternatives, controls early writes. **Hypothesis retention:** without the buffer, initial holding remains possible—Gemini's hold frequency varies across runs, Qwen's is stable—but no held-event resolutions are ever recorded; the buffer keeps an addressable event and its alternatives available for later revision. **Revision integration:** disabling it

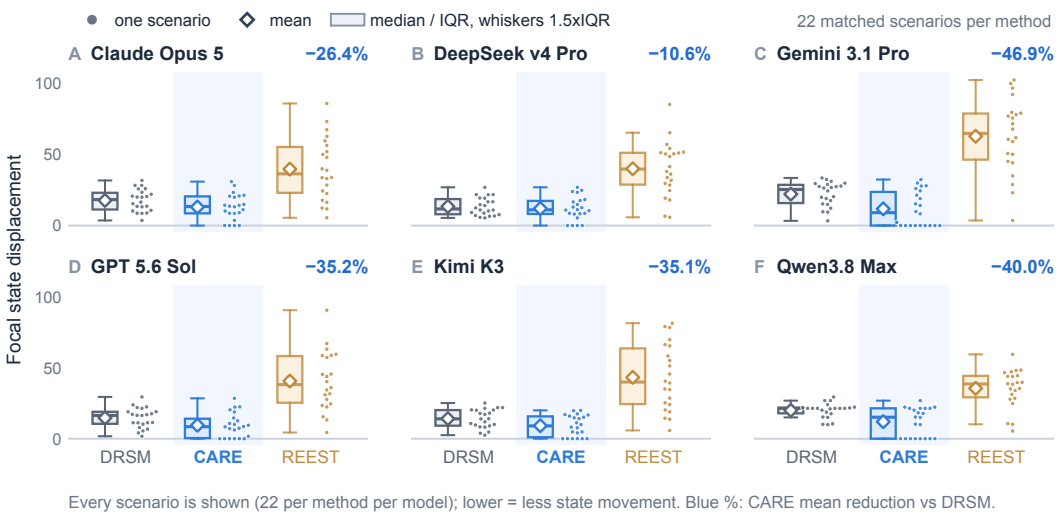

Figure 3: Historical focal state displacement under ambiguous evidence. Each panel is one model family; each method shows a median/IQR box with 1.5xIQR whiskers beside all 22 matched scenario values; diamonds mark means. Lower values mean less state movement.

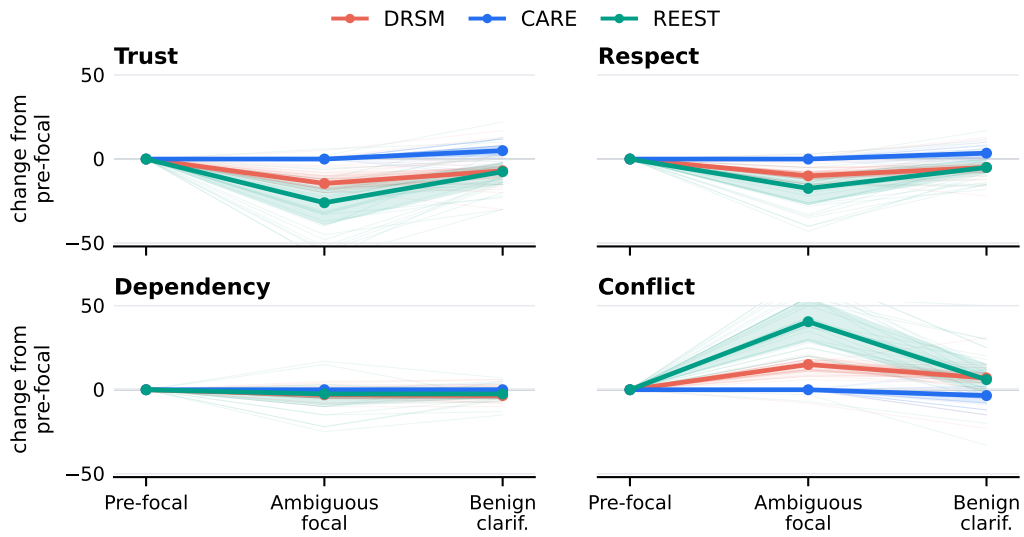

Figure 4: **Held focal events keep the state open for later revision.** Historical trajectories for the 32 matched focal-hold cases, centered on each method's pre-focal state (turns 2, 3, 10). Thin lines are individual cases; bold lines and bands are medians and IQRs. CARE moves only after benign clarification; DRSM and REEST drift early and revert.

leaves initial holding intact and lowers residual displacement, but direction scores fall from 0.86 to 0.35 (Gemini) and from 0.74 to 0.42 (Qwen): without revision, later evidence no longer steers the state.

4.4 Revision after clarification in paired worlds

On the benign branch, CARE's post-clarification AUEC is slightly lower than DRSM's; on the malicious branch it is substantially higher, i.e., farther from the reference intervals; per family, all ten malicious families favor DRSM while the benign families split five and five (Figure 5).

This evaluation separates the presence of a linked revision operator from the adequacy of its updates: CARE implements revision, yet its malicious-branch state remains farther from the reference range.

Branch	DRSM	CARE
Benign clarification	0.0146	**0.0134**
Malicious clarification	**0.0169**	0.0444
Overall DRSM − CARE	−0.0132	
(95% CI)	[−0.0258, 0.0030]	

Per family: all ten malicious families favor DRSM; benign splits five–five. Lower AUEC is closer to the branch reference; bold marks the better method per branch.

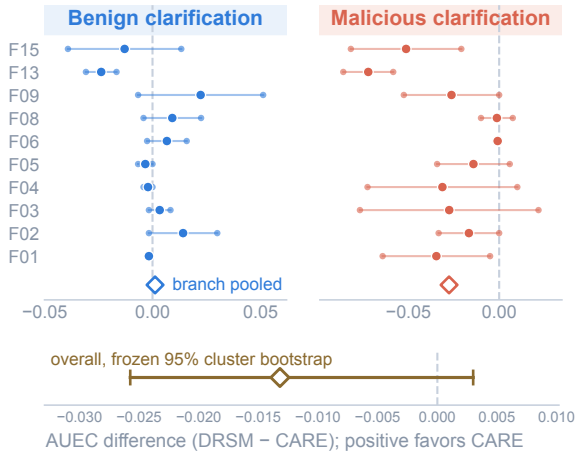

Figure 5: **Paired-world revision: branch outcomes and per-family contrasts.** Left: normalized post-clarification AUEC by clarification branch (lower is better), the overall DRSM-minus-CARE contrast with its two-way cluster-bootstrap 95% interval, and the per-family split. Right: per-family paired differences (DRSM − CARE; positive favors CARE); end dots are the two model families, filled dots their means, open diamonds branch-pooled values; the bottom strip repeats the overall contrast.

The overall DRSM-minus-CARE contrast is negative, but its cluster-aware 95% interval includes zero (Figure 5, left), so the benign-side difference does not establish an overall advantage. Together with the component analysis, the comparison identifies two distinct design requirements: **control premature writes** and **make later revisions adequately responsive** to the resolved evidence. Robustness checks are in Appendix F.

5 Conclusion

We introduced **CARE** to address *premature commitment* in relational memory. Its state-commitment contract separates effects assessed as shared across candidate interpretations from effects tied to a particular reading, implemented by **Consensus Write**, **Hypothesis Buffer**, and **Evidence Replay** through selective writing, retention of unresolved effects, and revision linked to their source events. Historical diagnostics show lower focal displacement than immediate updating across six model families; ablations separate the roles of gating, retention, and revision; and the paired-world diagnostic scores later revision apart from initial writing. CARE thus makes **selective commitment with a path to revision** an explicit memory operation, implementing the principle ***remember the evidence without prematurely settling its meaning***.

6 Limitations

Focal displacement is a method-neutral diagnostic of state movement at unresolved events, not a correctness measure: **less state movement does not by itself establish greater accuracy,** and the premature-commitment reading rests on controlled comparisons and ablations, not displacement magnitudes alone. CARE's contract likewise controls when assessed effects enter state but does not guarantee that the upstream interpretations or effects are correct: it depends on candidate interpretations, uncalibrated support scores, and assessed effects, and the gate cannot repair omitted readings or mistaken assessments; focal-event identifiers and clarification links are supplied, not discovered.

The paired-world evaluation separates early commitment from later revision but spans two model families and ten scenario families, scores states only at clarification and the next event, and uses LLM-adjudicated reference intervals (Appendix E.1), not measurements of real relationships. It shows that **avoiding premature commitment does not by itself guarantee adequate later revision**— CARE's malicious-branch state remains farther from the reference—yet the source of this revision gap is not isolated. Longer-horizon evidence comes from the historical implementation, so it is not a full long-horizon validation of the revised rules.

AI USE STATEMENT

In this work, Generative AI tools were used to generate synthetic datasets, dialogues, and experimental scenarios; propose and refine research hypotheses; translate manuscript and research materials; support qualitative and thematic analysis; and assist in interpreting experimental results. The authors reviewed all AI-assisted work, verified the resulting materials and interpretations, and take full responsibility for the final content, claims, and artifacts presented in this paper.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1573–1596. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.116. URL https://aclanthology.org/2023.eacl-main.116/.

Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327, 2005. doi: 10.1145/1067860.1067867.

Timothy W. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Timothy Heeren. It's just like you talk to a friend: Relational agents for older adults. *Interacting with Computers*, 17(6):711–735, 2005. doi: 10.1016/j.intcom.2005.09.002. URL https://doi.org/10.1016/j.intcom.2005.09.002.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1470.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-acl.125.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 610–614. European Language Resources Association, 2020.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.acl-long.847.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11370–11403. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.699.

Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89 (1–2):1–29, 1997. doi: 10.1016/S0004-3702(96)00038-0.

Pengfei Du. Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670*, 2026.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z. Pan. Rethinking memory in AI: Taxonomy, operations, topics, and future directions. *arXiv preprint arXiv:2505.00675*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 653–670. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.48.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-0595.

Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, 1988.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5010–5028. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.344.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170, 1988. doi: 10.1007/BF00247909.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Bernal Jim'enez Guti'errez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1902.

Fritz Heider. *The Psychology of Interpersonal Relations*. John Wiley & Sons, 1958.

Robert A. Hinde. *Towards Understanding Relationships*. Academic Press, London, 1979.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of AI agents. *arXiv preprint arXiv:2512.13564*, 2025.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13584–13606, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.838. URL https://aclanthology.org/2023.emnlp-main.838/.

Yuru Jiang, Yang Xu, Yuhang Zhan, Weikai He, Yilin Wang, Zixuan Xi, Meiyun Wang, Xinyu Li, Yu Li, and Yanchao Yu. The CRECIL corpus: A new dataset for extraction of relations between characters in chinese multi-party dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2337–2344, Marseille, France, 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.250/.

Edward E. Jones and Keith E. Davis. From acts to dispositions: The attribution process in person perception. In Leonard Berkowitz (ed.), *Advances in Experimental Social Psychology*, volume 2, pp. 219–266. Academic Press, New York, 1965. doi: 10.1016/S0065-2601(08)60107-0.

Edward E. Jones and Victor A. Harris. The attribution of attitudes. *Journal of Experimental Social Psychology*, 3(1):1–24, 1967. doi: 10.1016/0022-1031(67)90034-0.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems*, 2024.

Hirofumi Katsuno and Alberto O. Mendelzon. On the difference between updating a knowledge base and revising it. In James F. Allen, Richard Fikes, and Erik Sandewall (eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pp. 387–394. Morgan Kaufmann, 1991. URL https://kr.org/proceedings/KR-1991-proceedings-scanned.pdf.

Harold H. Kelley. The processes of causal attribution. *American Psychologist*, 28(2):107–128, 1973. doi: 10.1037/h0034225.

Harold H. Kelley, Ellen Berscheid, Andrew Christensen, John H. Harvey, Ted L. Huston, George Levinger, Evie McClintock, Letitia Anne Peplau, and Donald R. Peterson. *Close Relationships*. W. H. Freeman, New York, 1983.

Eunsu Kim, Junyeong Park, Juhyun Oh, Kiwoong Park, Seyoung Song, A. Seza Doğruöz, Alice Oh, and Najoung Kim. Are they lovers or friends? evaluating LLMs' social reasoning in english and korean dialogues. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23431–23451, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1074. URL https://aclanthology.org/2026.acl-long.1074/.

Eunwon Kim, Chanho Park, and Buru Chang. SHARE: Shared memory-aware open-domain long-term dialogue dataset constructed from movie script. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14474–14498. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.704.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12930–12949. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.799.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 1378–1387. PMLR, 2016.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. *arXiv preprint arXiv:2605.05583*, 2026.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pp. 15–26. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.conll-1.2.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24868–24891, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1264. URL https://aclanthology.org/2025.emnlp-main.1264/.

Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995. doi: 10.5465/amr.1995.9508080335.

Daniel J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38(1):24–59, 1995. doi: 10.2307/256727.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1400–1409. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1147.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4975–5001, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.257. URL https://aclanthology.org/2025.naacl-long.257/.

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*. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1050.

Liang Qiu, Yuan Liang, Yizhou Zhao, Pan Lu, Baolin Peng, Zhou Yu, Ying Nian Wu, and Song-Chun Zhu. SocAoG: Incremental graph parsing for social relation inference in dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 658–670. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.54.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1534.

John K. Rempel, John G. Holmes, and Mark P. Zanna. Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1):95–112, 1985. doi: 10.1037/0022-3514.49.1.95.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15198–15216, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.940. URL https://aclanthology.org/2023.emnlp-main.940/.

Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404, 1998. doi: 10.5465/amr.1998.926617.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *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. 4463–4473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

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*, 2023.

Joykirat Singh, Zaid Khan, Archiki Prasad, Justin Chih-Yao Chen, Akshay Nambi, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Agent-BRACE: Decoupling beliefs from actions in long-horizon tasks via verbalized state uncertainty. *arXiv preprint arXiv:2605.11436*, 2026.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071.

Zikai Song, Xiajie Li, Yunyao Zhang, Xinglang Zhang, Wei Yang, and Junqing Yu. Social intelligence modeling: A comprehensive survey from social perception to social simulation. ResearchGate preprint, 2026. URL https://www.researchgate.net/publication/411123257_Social_Intelligence_Modeling_A_Comprehensive_Survey_from_Social_Perception_to_Social_Simulation. Preprint available on ResearchGate.

Wolfgang Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In *Causation in Decision, Belief Change, and Statistics*, volume II, pp. 105–134. Springer, 1988.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.989.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.413.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. PRIDE: Predicting relationships in conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4636–4650. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.380.

Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. Towards objectively benchmarking social intelligence of language agents at the action level. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8885–8897, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.526. URL https://aclanthology.org/2024.findings-acl.526/.

Bowen Wu, Wenqing Wang, Haoran Li, Yunhan Deng, Ying Li, Jingsong Yu, and Baoxun Wang. Interpersonal memory matters: A new task for proactive dialogue utilizing conversational history. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 47–67. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.conll-1.4.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025b.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to AI memory: A survey on memory mechanisms in the era of LLMs. *arXiv preprint arXiv:2504.15965*, 2025c.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.207.

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 *International Conference on Learning Representations*, 2023.

Keen You and Dan Goldwasser. "where is this relationship going?": Understanding relationship trajectories in narrative text. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 168–178, Barcelona, Spain, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.starsem-1.18/.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4927–4940. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.444.

Amir Zadeh, Minghai Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019. doi: 10.1109/CVPR.2019.00901.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1496.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024.