

Whose Memory Is It? Scope-Aware Commit Rules for Long-Term LLM Memory

Hongyu Gu¹

University of Science
and Technology of China
Hefei, China

ustc_23ghy@mail.ustc.edu.cn

Xinchang Li²

University of Science
and Technology of China
Hefei, China

secirst@mail.ustc.edu.cn

Abstract

Persistent memory allows an LLM agent to carry experience across conversations, but it also turns a local reasoning mistake into a durable one. During deliberation, an agent may consider a plan, simulate a tool result, report another speaker’s belief, and then reject all of them. If memory retains only the resulting sentences, those once-useful possibilities can later return as facts. The record is neither fabricated nor irrelevant; it has simply been detached from the context in which it was valid. We identify this missing context as *discourse ownership*: the world, branch, or speaker that licenses a proposition. Our first finding is counterintuitive. Language models already carry a causally active signal for ownership, yet conventional memory interfaces discard it when they convert reasoning into records. We introduce CASK (Causally Anchored Scoping Keys), a commit rule that preserves this signal so that shared-world facts enter durable memory while provisional content remains available only within its original scope. Our second finding is that the most obvious way to preserve the signal—storing the discovered internal coordinates—is unreliable because equivalent representations need not keep the same coordinates. CASK instead preserves the stable relations that express ownership. Controlled long-conversation conflicts and tool-agent traces show that this design improves memory admission and prevents provisional content from contaminating later answers while complementing runtime provenance. The resulting commit boundary lets agents explore more possibilities without granting every intermediate sentence authority over future behavior.

1 Introduction

An agent spends ten minutes considering Paris, rejects the trip, and chooses Sydney. Months later, its memory retrieves “move to Paris” as a highly relevant personal fact. Nothing was hallucinated: the model really did produce that sentence. The error arose because the system remembered the proposition and forgot the world in which the proposition was licensed.

This failure grows out of progress in reasoning. Chain-of-thought and reinforcement-learned reasoning extend deliberation; Tree of Thoughts explores competing branches; ReAct interleaves plans with actions; Reflexion retains failed attempts; and latent-reasoning methods carry possible continuations before committing to one [Wei et al., 2022, Yao et al., 2023a,b, Shinn et al., 2023, Hao et al., 2024, DeepSeek-AI, 2025]. Each technique gains power by producing content whose status is provisional: a belief under revision, an unexecuted plan, a hypothetical tool outcome, or a discarded branch. More test-time computation can improve the final answer while creating more text that must never become an unconditional premise. Long context postpones this decision; persistent memory makes it consequential across episodes.

Linguistic theory supplies the missing distinction. “The appointment is Friday,” “Mina believes the appointment is Friday,” and “suppose the appointment were Friday” can express the same proposition while licensing different discourse updates. Mental-space accounts track the context—shared reality, a belief, or a hypothetical—that owns a proposition [Fauconnier, 1994, Kamp and Reyle, 1993, Heim, 1982, Roberts, 1996]. Most retrieval systems instead compress a record into content and ask whether it is relevant [Lewis et al., 2020, Maharana et al., 2024, Packer et al., 2023]. Relevance is necessary, but it is not permission: the most relevant record may still belong only to a belief, simulation, or abandoned plan.

This perspective changes the diagnosis. The dangerous record need not be false, irrelevant, or poorly generated. It may be perfectly faithful to a reasoning branch and still be unsafe as a premise about the shared world. A memory system must therefore recover an attachment—which discourse owner licenses the proposition—rather than assign the proposition a context-free truth score. This is the commit problem studied in this paper.

Our first finding is counterintuitive: the model already represents the distinction that its memory layer forgets. Interchange interventions redirect which mental space the model uses while preserving the proposition under discussion. Intervening on the same route also changes a frozen write decision in the predicted direction, whereas matched controls do not. The bottleneck is thus not simply insufficient model knowledge. It is a model–memory interface that exports content without exporting ownership.

Making this internal distinction durable reveals a second surprise. The coordinates of a discovered neural route are not a stable identity for what that route means: an equally valid refit can rotate the basis and change every coordinate without changing the represented distinction. Internal probes can absorb that symmetry; a long-lived memory database cannot. A persistent key should therefore preserve the relations that define the ownership signal, not the arbitrary axes used to express it.

We introduce CASK (Causally Anchored Scoping Keys), illustrated in Figure 1. CASK turns the model’s ownership-sensitive response pattern into a stable memory edge: actual-world content may enter shared memory, while plans, beliefs, and counterfactuals remain available under their original scope. Its relational key is unchanged by equivalent reorientations of the internal basis, yet retains the information needed for attachment. The resulting interface is small enough to sit at the commit boundary between reasoning and retrieval, and its theoretical guarantees explain why stored keys remain meaningful across equivalent representations.

The evaluation follows a record from internal routing through admission, retrieval, and final answering. CASK leads the evaluated attachment representations at both model scales, and its advantage concentrates on changes in reporting scope rather than payload. Direct interventions connect that gain to the ownership route; long-conversation conflicts show that cleaner attachment reduces later premise contamination; and generated tool-agent traces show how CASK works with execution metadata. The result changes how reasoning and memory can be co-designed: reasoning research gains a principled boundary between exploring a candidate and promoting it to durable state, agent builders gain a testable commit layer, and memory evaluation gains premise licensing as an objective alongside retrieval and question answering.

Our contributions are:

1. We identify discourse ownership as the missing commit type between contemporary reasoning and persistent memory, and show that the model’s ownership route causally controls a frozen write decision.
2. We develop a stable relational key that preserves this route across equivalent internal bases, with formal invariance, completeness, and perturbation guarantees.
3. We show how attachment shapes retrieval and answering across controlled conflicts and tool-agent traces, yielding a provenance-aware deployment policy and a premise-licensing evaluation target.

2 Related work

LoCoMo made long conversational histories measurable and motivates our use of naturally authored conversations [Maharana et al., 2024]. Retrieval and memory architectures optimize relevance, compression, or retention through memory tiers, reflections, and selective updates [Lewis et al., 2020,

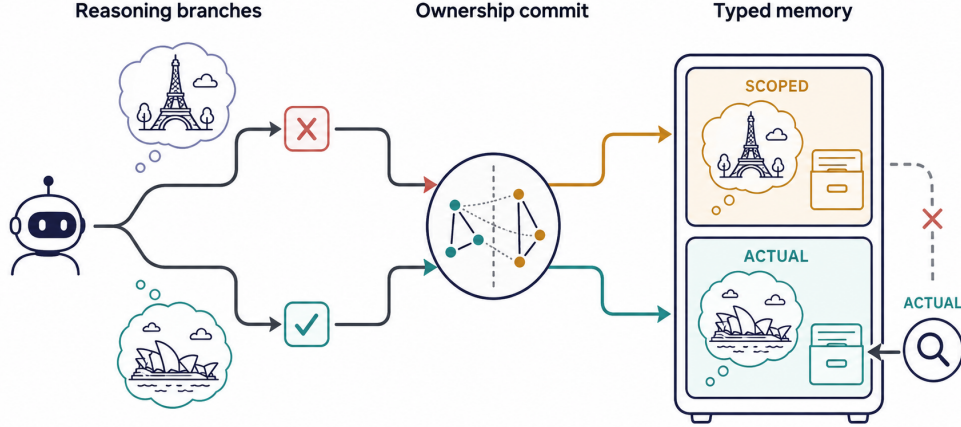


Figure 1: Reasoning creates propositions that belong to different worlds. The agent considers Paris but rejects that branch, while committing to Sydney. CASK converts the model’s ownership-sensitive relational geometry into an explicit commit edge: Sydney enters actual-world memory, whereas Paris remains available only inside its scoped world. A later query may use committed evidence without turning the discarded branch into a fact.

Packer et al., 2023, Park et al., 2023, Xiong et al., 2025]. They decide what persists and returns; CASK additionally records the discourse world in which it is licensed as a premise.

Mental spaces and dynamic semantics treat attitudes and modality as structured context change: identical content licenses different updates under a belief, assertion, or hypothetical [Fauconnier, 1994, Kamp and Reyle, 1993, Heim, 1982]. CASK instantiates this principle at the model–system boundary by storing a proposition with its attachment. Unlike truth scoring, it can preserve a faithful belief or counterfactual without letting it update the actual world.

Speaker-commitment and event-factuality work separates content, holder, and commitment status [de Marneffe et al., 2019, Jiang and de Marneffe, 2019, Murzaku et al., 2023, Li et al., 2024, Rovera et al., 2025]. Our question begins where tagging ends: how can a reasoning system preserve this distinction in persistent state?

Mechanistically, an *interchange intervention* replaces part of a target example’s hidden state with the corresponding state from a donor example and tests whether the target adopts the donor’s high-level variable. Distributed alignment search (DAS) learns the subspace on which to perform that replacement [Geiger et al., 2024, 2025, Wu et al., 2022]. We use intervention to validate routing, then separate that causal test from downstream prediction. Representation similarity work recognizes that equivalent subspace bases may differ by an orthogonal rotation [Kornblith et al., 2019]. A Gram matrix records all pairwise inner products among an ordered set of vectors; we use its classical completeness to preserve an ownership-responsive/complementary partition in a compact memory key.

Chain-of-thought, tree search, tool use, reflection, and latent reasoning generate candidate trajectories in different ways [Wei et al., 2022, Yao et al., 2023a,b, Shinn et al., 2023, Hao et al., 2024]. Once reused across episodes, they share one question: which intermediate conclusions were promoted to the agent’s shared world? CASK types that commit, complementing work that scales reasoning computation [DeepSeek-AI, 2025].

Existing frameworks usually resolve branches within one episode. Persistent agents add a second time scale: summaries and traces can return after branch control has disappeared, making a rejected path resemble an observation. An explicit commit operation joins within-episode search to cross-episode state.

3 Problem: memory requires discourse ownership

Let $c = (x, m, s)$ denote a candidate record: x is the preceding conversation, m the proposition proposed for storage, and s its reporting frame (assertion, belief, plan, fiction, or counterfactual). Its *discourse owner* $\omega(c)$ is the context that licenses m . Owners form $\Omega = \{\omega_0\} \cup \Omega_{\text{scoped}}$, where ω_0 is the shared actual world and Ω_{scoped} contains hypothetical worlds and *attitude holders*, the speakers or agents whose beliefs and plans supply a context. The binary target

$$a(c) = \mathbb{I}[\omega(c) = \omega_0] \quad (1)$$

uses the indicator $\mathbb{I}[\cdot]$ (one when its condition holds and zero otherwise) to state whether m may become an unconditional premise in shared memory. Thus *nonactual* means scoped, not false: such a record remains useful when a query addresses its owner.

The target governs the embedded proposition, not whether the episode is erased. If an agent considered moving to Paris and then chose Sydney, $\text{MOVETo}(\text{Paris})$ remains attached to the plan space, while the meta-event $\text{CONSIDERED}(\text{agent}, \text{MOVETo}(\text{Paris}), t)$ is itself an actual-world fact. A query about considered options may retrieve the latter; a query about the eventual destination may not use the embedded Paris proposition as an unconditional premise. The binary edge is therefore the minimal distinction needed for premise licensing, not a delete/retain decision.

The learned gate $g_\theta(c) \in [0, 1]$, with classifier parameters θ , produces $\hat{a}(c) = \mathbb{I}[g_\theta(c) \geq \tau]$ at threshold τ . A positive decision links m to ω_0 ; otherwise it retains its scoped owner. For query q , $A(q) \subseteq \Omega$ denotes the owners permitted by the query and $\rho(q, m) \in \mathbb{R}$ its relevance score. Retrieval ranks c by $\rho(q, m)$ subject to $\omega(c) \in A(q)$. An *attachment error* is $\hat{a}(c) \neq a(c)$; *contamination* occurs when a nonactual record enters an actual-world result.

We evaluate the same representational route at two levels. A causal-routing study asks whether interchange interventions control the model’s mental-space decision. The downstream study then asks whether a stable interface built from that ownership-responsive route improves memory admission. Together they test both the internal linguistic variable and its value after it crosses the model–memory boundary.

4 Method

4.1 Causal anchor and diagnostic configuration

DAS learns a low-dimensional subspace whose interchange interventions realize a hypothesized high-level variable [Geiger et al., 2024, 2025]. In the routing study, we train a rank-16 ownership subspace only on training constructions. A held-out *target* is the example being evaluated; its *donor* is a matched example with the opposite owner whose subspace state is inserted into the target. Success means that the target routes to the donor owner without copying the donor proposition.

For attachment, we reuse the fixed rank- $r = 16$ orthogonal projector $P = UU^\top \in \mathbb{R}^{d \times d}$, where $U \in \mathbb{R}^{d \times r}$ has orthonormal columns and d is the model’s hidden width. Let $U_\perp \in \mathbb{R}^{d \times (d-r)}$ be an orthonormal basis for the complementary subspace, so $UU^\top + U_\perp U_\perp^\top = I_d$. For each candidate record, the frozen model receives $k = 6$ fixed Yes/No diagnostic prompts covering shared-memory admission, premise licensing, speaker commitment, future factual use, actual-world ownership, and frame removal. If $h_i \in \mathbb{R}^d$ is the last-token hidden state after the selected transformer block for prompt i , their fixed prompt order gives $H = [h_1, \dots, h_k] \in \mathbb{R}^{d \times k}$. The first four prompts also give $\ell_i = z_{\text{Yes},i} - z_{\text{No},i}$, the difference between the two pre-softmax output scores, collected as $\ell \in \mathbb{R}^4$. Appendix B gives every prompt and extraction choice. We form

$$V_\parallel = U^\top H, \quad G_\parallel = V_\parallel^\top V_\parallel, \quad (2)$$

$$V_\perp = U_\perp^\top H, \quad G_\perp = V_\perp^\top V_\perp. \quad (3)$$

For symmetric $A \in \mathbb{R}^{k \times k}$, let $\text{vech}(A)$ stack its upper triangle including the diagonal, and let $\parallel$ denote vector concatenation. The composite feature is

$$\phi_{\text{CASK}} = \text{vech}(G_\parallel) \parallel \text{vech}(G_\perp) \parallel \ell \in \mathbb{R}^{46}, \quad (4)$$

Each 6×6 Gram block contributes 21 entries. After coordinate-wise standardization, a logistic-regression classifier with bias therefore has 47 learned parameters. The ownership-block-only

interface has 21 features and 22 classifier parameters; a mean full-hidden linear probe has 2,049. Throughout, $O(p) = \{Q \in \mathbb{R}^{p \times p} : Q^\top Q = I_p\}$ is the orthogonal group, $\|\cdot\|_2$ is the vector Euclidean or matrix spectral norm, and $\|\cdot\|_F$ is the Frobenius norm.

4.2 Why two Gram blocks?

The full Gram matrix satisfies $H^\top H = G_\parallel + G_\perp$. It is invariant but erases which relations belong to the ownership-responsive partition. CASK retains the two blocks separately, producing an invariant *typed decomposition*. The low-rank block supplies the causal anchor: interchange interventions establish that changing this route changes the model’s owner selection. The complement preserves the surrounding relational configuration, while output logits add the model’s explicit judgment. Together, the typed blocks retain the causal routing geometry and the context that makes a memory-commit decision identifiable.

We use one CASK representation and scorer for two policies. The default sets $\tau = 0.5$ for balanced attachment. CASK@R95 selects the inverse ℓ_2 -regularization strength C and τ on development conversations to maximize nonactual rejection subject to admitting at least 95% of actual records.

4.3 From a reasoning branch to a memory commit

The interface sits between deliberation and persistence. CASK keeps each candidate paired with its diagnostic configuration, separates ownership-responsive from complementary relations, and admits it to shared retrieval only after predicting an actual-world edge. Otherwise, the content remains available under its scoped owner rather than being deleted.

The causal projector ties each edge to an internal variable that controls the model’s routing behavior; the Gram construction makes that evidence stable across equivalent bases. This turns a mechanistic finding into a durable systems interface.

At retrieval time, all records receive the same relevance score. The query selects an admissible owner set, and the attachment edge masks records outside that set before the *reader*, the answer-generating model, sees them. Semantic retrieval answers *which content is useful*; the scope key answers *in which world that content may serve as a premise*. Experiments compute one BM25 lexical-relevance ranking and change only the admission mask.

Runtime provenance and CASK operate at complementary levels. A ledger supplies event status, while CASK supplies proposition-level scope within language, summaries, and reflections. Together they support a commit layer that carries execution evidence and linguistic ownership into persistent memory.

4.4 Formal guarantees

The following results state what the stored key preserves. A *gauge change* here means an orthogonal change of basis inside either fixed subspace; it changes coordinates but not the represented subspace or its inner products.

Theorem 1 (Gauge invariance). *For any $Q_\parallel \in O(r)$ and $Q_\perp \in O(d - r)$, replacing $(V_\parallel, V_\perp)$ by $(Q_\parallel V_\parallel, Q_\perp V_\perp)$ while holding ℓ fixed leaves ϕ_{CASK} unchanged.*

Remark 1 (Persistent interoperability). Theorem 1 keeps stored keys comparable when the same ownership subspace is exported in different orthonormal bases, as can occur across libraries, machines, or refactorings.

Theorem 2 (Completeness for ordered configurations). *For $V, W \in \mathbb{R}^{r \times k}$, $V^\top V = W^\top W$ if and only if there exists $Q \in O(r)$ such that $W = QV$. This holds even when the configurations are rank deficient.*

Remark 2 (No avoidable information loss). Theorem 2 makes the Gram key a *maximal invariant*: it retains all information except the unavoidable choice of orthogonal basis. Thus two ordered configurations collide exactly when that basis symmetry makes them equivalent.

Theorem 3 (Perturbation stability). *For positive integers p, k, n , let $V, E \in \mathbb{R}^{p \times k}$ and $\tilde{V} = V + E$. Then*

$$\|\tilde{V}^\top \tilde{V} - V^\top V\|_F \leq 2\|V\|_2\|E\|_F + \|E\|_F^2. \quad (5)$$

Moreover, for $\phi, \tilde{\phi}, w \in \mathbb{R}^n$, $b, t \in \mathbb{R}$, and the affine logit $f(x) = w^\top x + b$, the decision $\mathbb{I}[f(\phi) \geq t]$ is unchanged whenever $|f(\phi) - t| > \|w\|_2 \|\tilde{\phi} - \phi\|_2$.

Remark 3 (A commit certificate). Theorem 3 turns extraction drift into a test: an edge whose margin to its decision threshold exceeds the feature-error budget retains its admission decision.

Proposition 1 (Typed refinement). *The ordered pair $(G_\parallel, G_\perp)$ determines the full Gram matrix $H^\top H$, whereas $H^\top H$ alone, without knowing which rank- r projector is designated ownership-responsive, does not in general determine either block.*

Remark 4 (Why two blocks matter). Proposition 1 formalizes the ablation contrast: a full Gram key can preserve overall geometry while erasing which relations came from the causally identified ownership subspace. The typed pair preserves that distinction.

Proof sketches. Orthogonality gives $(QV)^\top(QV) = V^\top V$, proving invariance. Equal Gram matrices define an isometry between the column spans of two configurations, which extends to an orthogonal map and proves completeness. Expanding $(V + E)^\top(V + E) - V^\top V$ and applying the matrix-norm inequality gives stability. Finally, summing the typed blocks recovers the full Gram, whereas a two-dimensional counterexample shows that the sum cannot recover its allocation. Appendix A gives complete proofs.

5 Experimental design

Research questions. The evaluation follows the path from representation to deployment. RQ1 asks whether mental-space ownership is causally active in the frozen model. RQ2 asks whether intervening on that same route changes the frozen memory-commit decision. RQ3 tests whether a typed invariant export improves attachment and follows reporting scope rather than payload. RQ4 challenges the interface with unseen reporting frames, diagnostic paraphrases, and naturally generated tool-agent traces, including a provenance-first hybrid. RQ5 measures how the resulting edge changes retrieval integrity and final answering in long histories.

Backbone and data. Experiments use frozen Qwen2.5-3B-Instruct and Qwen2.5-1.5B-Instruct representations [Yang et al., 2024]. Block 20 and rank 16 on 3B were fixed in a pilot before held-out evaluation; the 1.5B replication matches relative depth at block 16 and transfers rank 16 without a new search. LoCoMo is a benchmark of long, multi-session conversations with 1,986 annotated questions [Maharana et al., 2024]; we use 486 questions with exact answer spans. A *source-isolated fold* groups by conversation and draws alternative answers only from its training conversations. Projector, regularization, and threshold choices use training/development data only.

Controlled conflicts. For each question we retain the *evidence payload*, the answer-bearing text proposed for storage, and place it in frames covering belief, plan, fiction, and counterfactual scope. Alternative answers come only from training conversations, with zero exact overlap against development or test answers. The answer-conflict study contains 972 candidate records for 486 questions. BM25 scores the common corpus once; each gate filters the same ranking. The reader then answers from the retrieved records; *conflicting-answer adoption* is the fraction of outputs that repeat the injected alternative answer.

Factorial isolation. Two paired datasets separate scope from payload. In *scope-only*, actual and nonactual examples contain byte-identical evidence payloads and differ only in reporting frame. In *payload-only*, both examples use the same actual reporting frame and differ only in the first answer span. Each variant contains 4,860 records (2,430 pairs), and a separate head measures the information available in each representation for each factor.

Baselines. The comparison isolates the source of attachment information. *Term frequency-inverse document frequency (TF-IDF)* tests lexical content. *Full hidden state* probes the mean 2,048-dimensional hidden vector; *basis coordinates* use its mean coordinates in the learned ownership basis. *Output logits* use the four explicit diagnostic judgments. *Causal Gram*, *complement Gram*, and their logit-augmented variants retain relations from one projector side. *Full Gram+logits* adds the two blocks before classification, erasing their type. *Random typed split+logits* substitutes a matched random rank-16 partition. *Typed split Gram (no logits)* removes explicit judgments from CASK,

Table 1: Frozen-gate mediator intervention on 3B. Δp is signed toward donor scope; intervals cluster by conversation.

Swapped state	Δp [95% CI]	Flip (%)
Ownership subspace, opposite scope	0.715 [0.656, 0.775]	75.8
Complement, opposite scope	0.048 [0.030, 0.069]	6.6
Random rank-16, norm matched	0.004 [0.003, 0.005]	0.6
Ownership subspace, same scope	-0.034 [-0.046, -0.021]	5.9

while *flat memory* admits every record without a scope gate. All classifiers share source-isolated folds and selection. We report attachment accuracy, recall among the first five retrieved records (recall@5), contamination, mean reciprocal rank (MRR), exact match (EM), token-level F1, and conflicting-answer adoption. Uncertainty uses 95% confidence intervals (CIs) obtained by resampling conversations.

Isolation and evaluation units. Every source-isolated fold builds its alternative-answer pool from training conversations only. Development data selects regularization and, for the CASK@R95 operating point, the threshold; test conversations remain untouched until final evaluation. Attachment is scored per record, retrieval per question, and reader outputs use one shared deterministic cache whenever their input memories are identical. Confidence intervals resample conversations rather than individual records, treating the ten long histories as the independent units. This alignment matters because a single conversation contributes many correlated questions and constructions.

6 Results

6.1 The routing variable is causally controllable

On Qwen2.5-3B, rank-16 interventions route 134 of 146 held-out cases to the target mental space (91.78%), while matched random subspaces reach 3.22%. Clustering the 74 target scenarios gives an intervention advantage of 88.56 points (95% CI [83.30, 93.34]). All scenarios preserve the target proposition, isolating a change in the owner used for reasoning. The independently fixed 1.5B replication shows 86.30% causal routing versus 20.86% for matched random subspaces, with donor content in 1.37% of cases. The cross-scale result establishes a reusable model-internal control variable for memory admission.

We next intervene on that route inside the six diagnostic passes and reevaluate the fitted gate. Replacing the ownership-subspace state with an opposite-scope donor moves commit probability by .715 toward the donor [95% conversation-clustered CI: .656, .775], succeeds in the intended direction for 97.5% of records, and flips 75.8% of decisions. The effect appears in both directions (.694 for actual→scoped; .736 for scoped→actual), while matched controls rarely flip the decision. Manipulating the ownership route therefore changes the frozen downstream write decision, connecting the linguistic variable directly to memory behavior.

Prompt and frame controls. The conversation-isolated gate reaches 90.94% under unseen meaning-preserving diagnostic paraphrases and 93.44% on unseen reporting frames, versus 72.81% for TF-IDF on the latter. Holding out both banks jointly yields 70.16% (chance 50%). The literal prompts, fold results, and unchanged-data reference (97.81%) appear in Appendix B.

6.2 A deployable invariant key leads at both scales

Table 2 gives the complete source-isolated comparison, from lexical and unrestricted hidden-state probes to invariant relational interfaces. CASK at $\tau = 0.5$ leads every evaluated representation on both backbones: 83.33% attachment on 3B and 79.82% on 1.5B. Complement Gram plus logits is second on 3B, 2.16 points behind; the unrestricted full-hidden probe is second on 1.5B, 2.35 points behind. Conversation-clustered comparisons on 3B exclude zero against both output logits and the random-decomposition control. The complete typed representation outperforms either block alone. This is the behavior predicted by the method: the causal block identifies the ownership-changing

Table 2: Complete source-isolated attachment comparison (%). All models use the same conversation folds and fold-local alternative sources. CASK uses its default symmetric threshold $\tau = 0.5$ and is best across every evaluated representation family on both frozen backbones.

Representation	Qwen2.5-3B	Qwen2.5-1.5B
TF-IDF	50.21	50.21
Basis coordinates	65.64	74.07
Full hidden state	72.12	77.47
Output logits	78.19	62.65
Causal Gram	61.42	71.71
Causal Gram + logits	79.53	70.06
Complement Gram	69.75	73.56
Complement Gram + logits	81.17	72.12
Random typed split + logits	78.81	73.35
Typed split Gram (no logits)	72.94	75.82
CASK ($\tau = 0.5$)	83.33	79.82

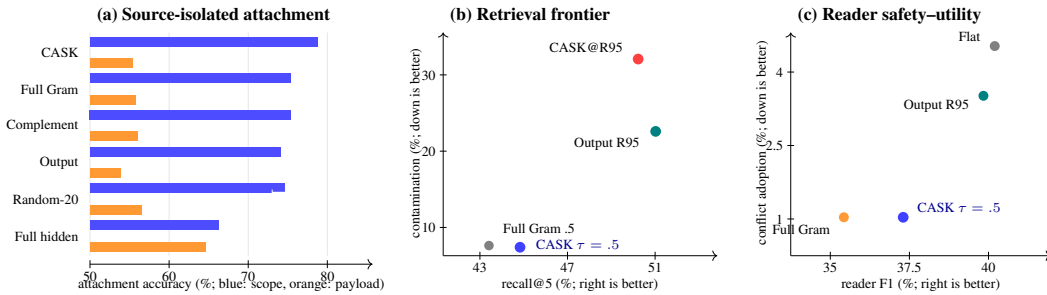


Figure 2: Controlled evidence for scope-sensitive memory. (a) The CASK advantage is concentrated in scope-only, not payload-only, attachment. Error bars on the random row denote one standard deviation across 20 decompositions. (b) At $\tau = 0.5$, CASK moves toward higher recall and lower contamination relative to full-Gram. Under a development-calibrated 95% actual-admission constraint, output logits define the stronger high-recall boundary. (c) The safety gain reduces adoption of the injected answer, while ordinary reader utility remains a separate trade-off.

route, the complementary block resolves its surrounding configuration, and preserving their relation produces the strongest commit rule.

The full-hidden result is especially informative: it stores 2,048 basis-dependent coordinates that change under equivalent reparameterizations, whereas CASK compresses the diagnostic response configuration into 46 basis-independent features while retaining the ownership-responsive/complementary split. The smaller typed relational representation is also the more accurate attachment interface.

The factorial control in Figure 2a reveals what the gain represents. When the evidence payload is held byte-identical and only its reporting scope changes, CASK reaches 78.70%, above every invariant control and all 20 random decompositions. Holding scope fixed removes the advantage. The gain is therefore specific to preserving who or which possible world licenses the content.

6.3 Effects propagate to conflicting memory

Table 3 compares CASK at $\tau = 0.5$ with the strongest invariant baseline on the answer-conflict study. CASK improves attachment by 2.05 points, increases recall@5 and MRR, and reduces contamination from 7.61% to 7.41%. Reader EM and F1 rise by 2.06 and 1.88 points, while injected-answer adoption ties at 1.03%. The downstream trajectory matches the statistically identified scope-only effect: preserving the typed partition improves admission first, then retrieval and answering.

Flat memory attains 29.63 EM and 40.19 F1 while adopting the conflict 4.53% of the time. Aggregate question answering (QA) and premise licensing are therefore distinct objectives: an unfiltered reader can answer more questions by accepting more unauthorized premises. CASK makes that trade-off measurable.

Table 3: Answer-conflicting memory. Attachment and retrieval/reader metrics are percentages except MRR. Lower is better for contamination and alternative-answer adoption.

Method	Att.	R@5	Cont.	MRR	EM	F1	Alt.↓
Full Gram+logits	81.28	43.42	7.61	31.08	24.69	35.43	1.03
CASK ($\tau = 0.5$)	83.33	44.86	7.41	32.35	26.75	37.31	1.03
Flat memory	–	–	–	–	29.63	40.19	4.53

6.4 The complete chain repeats at the smaller scale

The 1.5B model transfers rank 16 and fixes its layer by relative depth. Its intervention remains far above matched random routing, and its 79.82% attachment leads Table 2, 2.35 points above the 2,048-dimensional full-hidden probe despite using only 46 basis-independent features. At $\tau = 0.5$, CASK obtains 46.09% recall@5, 18.11% contamination, 26.13 EM, and 36.92 F1. The larger model has the cleaner safety profile; the smaller one shows that the ownership route and invariant export are already present below the principal scale.

6.5 One scorer supports safety-first and recall-first memory

The same score supports two policies without retraining. The symmetric $\tau = 0.5$ boundary gives the strongest balanced attachment and lowest internal-gate contamination. CASK@R95 instead selects a development threshold for at least 95% actual admission; it preserves flat-memory recall and reciprocal rank while reducing contamination by 11.19 points and conflict adoption by 1.17 points, with clustered intervals excluding zero. Its 29.42 EM and 40.30 F1 lead the evaluated internal-feature gates.

6.6 Natural agent traces favor a provenance-first hybrid

Across 75 generated tool-using episodes (511 records; ten conversations), runtime labels are hidden from the candidate text and folds remain source-isolated. A trace-adapted CASK fallback leads learned gates in balanced accuracy—the mean of actual-admission and scoped-rejection rates—at 91.10% versus TF-IDF’s 90.40%, and admits more actual observations (89.13% versus 81.52%). With exact provenance retained for half the records, the layered policy reaches 95.02%; a full ledger is the privileged upper bound. A 20-episode audit separates scoped propositions from their actual meta-events, and six-way batching costs 32.1 ms versus 24.8 ms for one diagnostic (1.29×). Full results are in Appendix B.

7 Implications and conclusion

More reasoning trajectories create more provisional propositions. CASK makes that growth compatible with persistent memory by requiring an ownership edge before a verbal, latent, or tool-using branch becomes a reusable premise. For reasoning researchers, it separates productive exploration from durable commitment; for agent builders, it provides a compact gate between deliberation and retrieval; for evaluation, it makes premise licensing measurable alongside relevance and QA.

The central innovation is to make linguistic ownership a first-class memory relation and ground that relation in a causally validated, basis-invariant neural representation. This gives the community a concrete way to test whether new reasoning methods remain reliable once their intermediate states outlive the episode that produced them.

Across two model scales, controlled conflicts, and generated traces, CASK leads every evaluated attachment representation, responds predictably to intervention, and works with runtime provenance. Extending ownership to nested scopes, learned retrievers, and broader model families is the next step toward memory systems that preserve not only more thought, but which world licenses each thought.

References

- Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung 23*, volume 23, pages 107–124, 2019. doi: 10.18148/sub/2019.v23i2.601. URL <https://doi.org/10.18148/sub/2019.v23i2.601>.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Gilles Fauconnier. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press, 1994.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR, 2024. URL <https://proceedings.mlr.press/v236/geiger24a.html>.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah D. Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.
- Irene Heim. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts Amherst, 1982.
- Nanjiang Jiang and Marie-Catherine de Marneffe. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1412. URL <https://aclanthology.org/P19-1412/>.
- Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, 1993.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of ICML*, 2019.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
- Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. MAVEN-FACT: A large-scale event factuality detection dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11140–11158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.651. URL <https://aclanthology.org/2024.findings-emnlp.651/>.
- 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)*, pages 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL <https://aclanthology.org/2024.acl-long.747/>.
- John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. Towards generative event factuality prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 701–715. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.44. URL <https://aclanthology.org/2023.findings-acl.44/>.

- 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 Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 2:1–2:22, 2023. doi: 10.1145/3586183.3606763.
- Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *OSU Working Papers in Linguistics*, 49:91–136, 1996.
- Marco Rovera, Serena Cristoforetti, and Sara Tonelli. Modafact: Multi-paradigm evaluation for joint event modality and factuality detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6378–6396. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.coling-main.425/>.
- 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*, volume 36, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.
- Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D. Goodman. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.318. URL <https://aclanthology.org/2022.naacl-main.318/>.
- Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025. URL <https://arxiv.org/abs/2505.16067>.
- An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL <https://arxiv.org/abs/2412.15115>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b.

A Additional proof details

Proof of Theorem 1. For either block V and orthogonal Q , $(QV)^\top(QV) = V^\top Q^\top QV = V^\top V$. Vectorization and concatenation with unchanged logits preserve equality.

Proof of Theorem 2. If $W = QV$ with $Q^\top Q = I_r$, then $W^\top W = V^\top V$. Conversely, assume $V^\top V = W^\top W$. Let $\text{col}(V) = \{V\alpha : \alpha \in \mathbb{R}^k\}$ and define $T : \text{col}(V) \rightarrow \text{col}(W)$ by $T(V\alpha) = W\alpha$. The map is well defined because $V\alpha = 0$ implies $\|W\alpha\|_2^2 = \alpha^\top W^\top W\alpha = 0$. For all $\alpha, \beta \in \mathbb{R}^k$, Gram equality gives $\langle T(V\alpha), T(V\beta) \rangle = \langle V\alpha, V\beta \rangle$, so T is an isometry onto $\text{col}(W)$. Completing orthonormal bases of the two column spans extends T to some $Q \in O(r)$, yielding $W = QV$. Equal Gram ranks make the same construction valid for rank-deficient configurations.

Proof of Theorem 3. Expand the Gram difference as $V^\top E + E^\top V + E^\top E$. The triangle inequality and $\|A^\top B\|_F \leq \|A\|_2 \|B\|_F$ give the stated bound. For the classifier, Cauchy–Schwarz gives $|f(\tilde{\phi}) - f(\phi)| \leq \|w\|_2 \|\tilde{\phi} - \phi\|_2$. If this change is smaller than $|f(\phi) - t|$, the two logits remain on the same side of threshold t .

Proof of Proposition 1. The forward map is addition: $H^\top H = G_\parallel + G_\perp$. For non-identifiability, take $d = 2$, $r = 1$, $k = 1$, and $H = (1, 0)^\top$. With ownership projector $P_1 = \text{diag}(1, 0)$, the typed pair is $(G_\parallel, G_\perp) = (1, 0)$. Under the equally ranked projector $P_2 = \text{diag}(0, 1)$ it is $(0, 1)$. Both yield the same stored full Gram $H^\top H = 1$, so that value alone cannot recover the typed allocation.

Corollary 1 (Composite commit stability). *Let $V_\parallel, E_\parallel \in \mathbb{R}^{r \times k}$, $V_\perp, E_\perp \in \mathbb{R}^{(d-r) \times k}$, and $\ell, e_\ell \in \mathbb{R}^4$. For $j \in \{\parallel, \perp\}$, set $\tilde{V}_j = V_j + E_j$ and $\tilde{\ell} = \ell + e_\ell$. Let $\phi = \phi_{\text{CASK}}$ from Equation (4), construct $\tilde{\phi}$ analogously from the perturbed quantities, and define $\Delta\phi = \tilde{\phi} - \phi$ and*

$$b_j = 2\|V_j\|_2 \|E_j\|_F + \|E_j\|_F^2.$$

Then the complete CASK feature obeys

$$\|\Delta\phi\|_2 \leq \sqrt{b_\parallel^2 + b_\perp^2 + \|e_\ell\|_2^2}. \quad (6)$$

Let $D \in \mathbb{R}^{46 \times 46}$ be the fixed positive diagonal training-set standardizer (constant coordinates use unit scale), let $\mu, \bar{w} \in \mathbb{R}^{46}$ and $\bar{b} \in \mathbb{R}$, and write the fitted logit as $F(\phi) = \bar{w}^\top D^{-1}(\phi - \mu) + \bar{b}$. Define $w_{\text{eff}} = D^{-\top} \bar{w}$ and $t_\tau = \log(\tau/(1 - \tau))$. The admission decision at any $\tau \in (0, 1)$ is preserved whenever $|F(\phi) - t_\tau|$ exceeds $\|w_{\text{eff}}\|_2$ times the right-hand side.

Proof. Apply Theorem 3 to each Gram block. Upper-triangle vectorization has Euclidean norm no larger than the Frobenius norm of the corresponding matrix, so the two block perturbations have norms at most $b_\parallel$ and $b_\perp$. Concatenation with the logit perturbation gives the displayed bound by the Pythagorean identity. Since $F(\tilde{\phi}) - F(\phi) = w_{\text{eff}}^\top \Delta\phi$, the decision claim follows from Theorem 3 with threshold t_τ . $\square$

Corollary 1 turns measured representation drift into a certificate for the deployed, standardized commit rule at either operating point.

B Exact diagnostic and selection protocol

The implemented interface uses one user message containing “Candidate memory:”, the complete framed record, and one of the following prompts. The Qwen chat template appends the generation prefix; no system message or decoded answer is used. We read the last-token output of the selected decoder block (the post-block residual returned by that block).

1. Should the embedded proposition enter unconditional shared memory? Answer only Yes or No.
2. Can this proposition be used as a premise about what actually happened? Answer exactly Yes or No.
3. Does the speaker commit this proposition to the common ground? Answer exactly Yes or No.
4. May a future agent retrieve this proposition as an established fact? Answer exactly Yes or No.
5. Is this proposition owned by the actual world rather than a hypothetical or attitude context? Answer exactly Yes or No.
6. Would storing this proposition without its reporting frame preserve its status? Answer exactly Yes or No.

All six ordered states enter both Gram matrices. Here z_t denotes the model’s pre-softmax score for token t ; the logit block is $(z_{\text{Yes}} - z_{\text{No}})$ for prompts 1–4, in that order, yielding $21 + 21 + 4 = 46$ features. We vectorize each symmetric Gram matrix by its upper triangle in row-major order: $(1, 1), (1, 2), \dots, (1, 6), (2, 2), \dots, (6, 6)$. Prompts are evaluated without sampling and are truncated only beyond 256 input tokens.

Block 20 and rank 16 were fixed in a pilot before the final held-out evaluation; the 1.5B study transferred rank 16 and matched relative depth at block 16. A development-only 3×3 grid over blocks $\{18, 20, 22\}$ and ranks $\{8, 16, 32\}$ maps sensitivity around that choice on one fixed conversation

split. Table 4 reports attachment accuracy and the aligned frozen-gate intervention effect. All nine settings retain 97.7–100% development accuracy and a positive causal effect (.305–.869), showing that the ownership route persists across nearby layers and ranks. Main test results use the prespecified block-20/rank-16 configuration.

Table 4: Development sensitivity. Each cell is attachment accuracy (%) / aligned Δp under the ownership-subspace swap.

	Rank 8	Rank 16	Rank 32
Block 18	99.2/.636	100.0/.305	98.4/.312
Block 20	98.4/.788	97.7/.716	99.2/.683
Block 22	100.0/.776	100.0/.801	100.0/.869

We retained each fold’s training conversations and replaced the prompt bank, the reporting-frame bank, or both at test time; projector and gate fitting use the original banks. The five test-fold accuracies were 84.38, 90.63, 92.97, 87.50, 99.22 for diagnostic paraphrases (mean 90.94); 91.41, 96.09, 87.50, 97.66, 94.53 for unseen reporting frames (mean 93.44); and 65.63, 67.97, 73.44, 68.75, 75.00 for the joint shift (mean 70.16). The unchanged-data reference is 97.81 and chance is 50%. Strong separate transfer shows that the ownership signal generalizes across prompt wording and reporting frames; the joint shift reveals their interaction.

Agent-trace protocol. We selected up to ten exact-span LoCoMo questions from each conversation, balanced across question categories where available, obtaining 75 eligible episodes. For each episode, Qwen2.5-3B-Instruct generated a tentative hypothesis, proposed search action, and unexecuted alternative; the runtime then executed conversation retrieval and returned either a successful observation or a failed-call event before the model produced a rejection and final conclusion. We removed event names and status fields before running the diagnostics. The runtime ledger labels successful observations actual; tentative, planned, alternative, rejected, and failed-call content scoped; and a final conclusion actual when supported by returned evidence. This produces 511 records: 92 actual and 419 scoped, grouped into five folds by source conversation. Each fold fits a trace-adapted head; the controlled-data head’s direct-transfer score (31.66% balanced accuracy) quantifies the domain shift.

Table 5: Generated tool-agent traces (%). BAcc is balanced accuracy; the 50% hybrid averages five metadata-retention masks.

Commit policy	BAcc	Actual	Scoped $\uparrow$
Output logits	82.29	81.52	83.05
Complement Gram+logits	86.37	85.87	86.87
TF-IDF	90.40	81.52	99.28
CASK fallback	91.10	89.13	93.08
CASK + 50% provenance	95.02	93.48	96.56
Runtime provenance (100%)	100.00	100.00	100.00

The provenance hybrid replaces the learned score with the runtime label wherever metadata is retained. At 0% retention it is exactly the CASK fallback; at 50% we average five independently sampled masks, obtaining 95.02% balanced accuracy, 93.48% actual admission, and 96.56% scoped rejection; at 100% it is the runtime-oracle upper bound. A per-episode meta-fact audit stores each scoped embedded proposition under an actual process wrapper (e.g., CONSIDERED) and verifies that process queries and world-state queries select the corresponding views.

Inference cost. Latency uses 200 records, 20 warm-up records, deterministic inference, and an A100-SXM4-80GB. At online batch size one, one diagnostic takes 24.83 ms median/26.42 ms p95 and 5.89 GiB peak memory. Six serial diagnostics take 148.76/153.76 ms, whereas batching the prompts takes 32.10/35.70 ms and 6.06 GiB. The resulting $1.29\times$ online multiplier requires no custom prefix-cache optimization.

C Reproducibility summary

All balanced learned systems use threshold 0.5. The source-isolated variants contain 4,860 records and 2,430 matched pairs each. Conversation clustering treats each of the ten conversations equally; paired bootstrap intervals resample conversations. Random-decomposition reporting uses one prespecified draw for clustered comparisons and 20 draws for the distributional control. The causal-routing test contains 146 held-out target cases grouped into 74 target scenarios. The answer-conflict task contains 972 records paired across 486 exact-span questions. Reader generations are deterministic for identical cached inputs, and model and threshold selection use training/development data.

Compute environment and existing assets. Experiments ran on an internal Linux server with NVIDIA A100-SXM4 80GB GPUs; GPU processes were single-device and at most two GPUs were active concurrently. The environment records Python 3.9, PyTorch 2.5.1, Transformers 4.55.0, NumPy 1.26.4, and scikit-learn 1.5.2. We obtained LoCoMo from its [official repository](#) under CC BY-NC 4.0. Qwen2.5-1.5B-Instruct is distributed under Apache 2.0, while Qwen2.5-3B-Instruct uses the Qwen Research License; both came from the official [Qwen organization](#). This non-commercial research use follows those terms, and the submission redistributes neither corpus nor model weights.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and Sections 1 and 7 state the contribution, evidence boundary, and non-universal scope explicitly.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Section 7 discusses dataset and model coverage, the binary ownership edge, compound distribution shift, trace calibration, retrieval choice, and deployment limits.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 4 states and proves three theorems and one proposition; Appendix A states and proves the composite stability corollary for the implemented feature.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 4–6 and Appendix B specify the representations, split isolation, sample counts, selection boundary, intervention controls, agent-trace construction, latency protocol, and clustered uncertainty used for the claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: LoCoMo and Qwen2.5 are public and credited, but no anonymous implementation or derived-data artifact accompanies this submission.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Sections 4–6 and Appendix B specify ranks, feature dimensions, grouped splitting, thresholds, development-only selection, agent-trace labels, and evaluation units needed to interpret the results.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: Section 6 reports conversation-clustered bootstrap intervals and distinguishes them from the standard deviation across 20 random decompositions.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not include experiments.
- The authors should answer [\[Yes\]](#) if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[No\]](#)

Justification: Appendix B reports the NVIDIA A100-SXM4 80GB hardware, single-device GPU processes, a two-GPU concurrency limit, and the software environment; exact per-run and aggregate wall-clock times were not retained.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The study uses released research assets, reports negative results and limitations, and makes no claims about protected attributes or human participants.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Sections 1 and 7 discuss preventing unauthorized memory promotion and the risks of incorrect attachment, false rejection, and overgeneralization.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not release a new pretrained model or high-risk scraped dataset.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
 - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
12. **Licenses for existing assets**
 Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
 Answer: [Yes]
 Justification: Appendix B credits the official sources and identifies LoCoMo as CC BY-NC 4.0, Qwen2.5-1.5B-Instruct as Apache 2.0, and Qwen2.5-3B-Instruct under the Qwen Research License; the study uses them for non-commercial research without redistributing the assets.
 Guidelines:
- The answer [N/A] means that the paper does not use existing assets.
 - The authors should cite the original paper that produced the code package or dataset.
 - The authors should state which version of the asset is used and, if possible, include a URL.
 - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
 - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
 - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
 - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
 - If this information is not available online, the authors are encouraged to reach out to the asset’s creators.
13. **New assets**
 Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
 Answer: [N/A]
 Justification: The study constructs evaluation records for its experiments but does not introduce a separately released dataset or model artifact.
 Guidelines:
- The answer [N/A] means that the paper does not release new assets.
 - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
 - The paper should discuss whether and how consent was obtained from people whose asset is used.
 - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
14. **Crowdsourcing and research with human subjects**
 Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
 Answer: [N/A]
 Justification: No crowdsourcing or new human-subject study was conducted.
 Guidelines:
- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
 - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
 - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.
15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
 Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals

(or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No new research with human participants was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Sections 4–6 identify the Qwen2.5-3B-Instruct and Qwen2.5-1.5B-Instruct backbones as frozen representation and diagnostic-response models and describe their role in the method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.