CAN AGENTS DEVELOP SHARED SHORTHAND WITH YOU? BENCHMARKING DYADIC CONVENTION FORMATION IN LONG-TERM CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term conversational agents must do more than recall facts: they should acquire the partner-specific expressions that emerge through repeated interaction. Existing long-term memory benchmarks test information stated in dialogue rather than conventions that emerge across sessions, leaving open whether an agent can participate in their formation and later use them. We introduce **PactBench**, a benchmark of 200 dyadic conventions spanning inversion, anchoring, adoption and erosion. Each convention has user-compression and agent-compression histories that test whether the agent can compress a reference, accept its partner's shortened form and later produce the established form. We evaluate each required capability independently by withholding the agent turn at that session while preserving the reference formation trajectory. To prevent plausible but ungrounded replies from passing, conjunctive scoring combines a direct meaning question with capability-specific checks wherever a form has been established. Experiments with four backbone LLMs and six existing memory systems identify two failures: (i) LLM-based memory systems lose or fail to retrieve the form–meaning link, and (ii) even when given the complete preceding reference formation history, agents do not reliably compress or produce the form. We present **Pact-Mem**, which preserves convention-level form–meaning bindings across sessions. Across four backbone LLMs, Pact-Mem consistently improves convention production over existing memory systems by up to 31 percentage points. Pact-Mem provides a convention-aware memory module for preserving and using dyadic conventions in long-term conversations.

1 INTRODUCTION

LLM-driven chat assistants, role-play characters and companions increasingly sustain multi-session conversations rather than answer isolated queries (Wu et al., 2025; Li et al., 2025a; Wang et al., 2024; Liu et al., 2026; Wu et al., 2026d;a). Their backbone LLM, long-term memory and agent framework determine how prior interaction shapes later replies (Chhikara et al., 2025; Li et al., 2025b; Rasmussen et al., 2025; Sun et al., 2025; Kang et al., 2026).

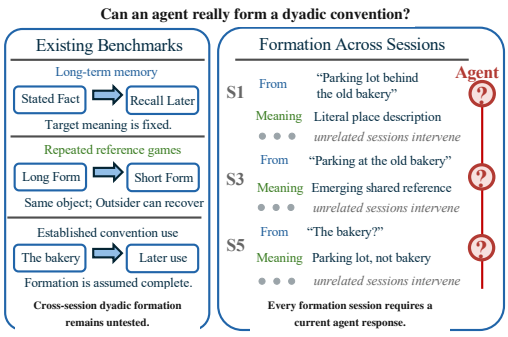

Figure 1: Existing evaluations assume a fixed target or completed formation. We test dyad-specific convention formation across interrupted sessions.

Two people in a long conversation settle on their own ways of referring to things. Psycholinguistics calls the result a *conceptual pact*, an agreement on how to refer to something that a pair settles through use rather than by stating it (Brennan & Clark, 1996). Repeated interaction between the same two speakers shortens such an expression (Clark & Wilkes-Gibbs, 1986), and the agreement binds to the pair that produced it rather than to the language (Brennan & Clark, 1996). We call these *dyadic conventions*. For example, two friends who park in the same spot behind an old bakery call it "the usual spot behind the old bakery" at first,

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.

and a few sessions later just "the bakery", dropped into a message with nothing attached. An outside reader cannot recover the dyad's meaning from the phrase alone without the dyad's prior interaction. Most long-term conversational-memory benchmarks do not capture this process. They test whether an agent can retrieve facts stated in conversation, including when the query expresses the fact indirectly (Li et al., 2026). However, the target meaning is introduced within a single session and remains fixed thereafter, so no expression forms for it across sessions. Repeated-reference-game studies instead find that LLM agents fail to refer to the same object with progressively shorter expressions (Hua & Artzi, 2024; Jones et al., 2026). The contracted expression still describes a publicly visible object, allowing an outsider to recover its meaning. A complementary line supplies expressions with novel interpretations, sometimes within a long conversation, and tests whether an LLM can apply them downstream (Patel et al., 2023). The mapping is given rather than formed through interaction. Neither setting evaluates whether an agent can form and later use a pair-specific expression whose meaning depends on interactions across sessions. Figure 1 summarizes this gap.

We therefore ask: *Can a long-term conversational agent take part in forming and using a dyadic convention with its user?*

In this paper, we introduce **PactBench** to localize failures during dyadic-convention formation. It covers 200 conventions spanning four formation mechanisms: inversion, anchoring, adoption and erosion. Following the compression and acceptance roles of Clark & Wilkes-Gibbs (1986), we synthesize two six-session reference formation trajectories for each convention that assign compression to either the agent or the user. This separation tests whether the agent can propose a compressed form and whether it can accept one introduced by its partner. We insert these histories into LoCoMo-Ultra, an extension of LoCoMo designed to mimic long-term conversation (Maharana et al., 2024).

To prevent an early failure from obscuring later behavior, we evaluate compression, acceptance and production independently at their respective sessions along each formation trajectory. For each evaluation, we provide the preceding reference formation sessions, withhold the current reference agent reply and ask the agent to answer. To prevent plausible but ungrounded replies from passing, conjunctive scoring requires both a *Meaning* check and the corresponding capability-specific check (Section 4.2). Experiments with four backbone LLMs and six existing memory systems identify two failures: (i) LLM-based memory systems lose or fail to retrieve the form–meaning link, and (ii) even when given the complete preceding reference formation history, agents do not reliably compress or produce the form (Section 5). To address the memory failure, we propose **Pact-Mem**, which stores partner-specific form–meaning mappings and retrieves them through form-free situational views (Section 6). Across all four backbone LLMs, Pact-Mem consistently improves convention production over existing memory systems, with a gain of up to 31 percentage points.

We make the following contributions: (1) We propose **PactBench**, a benchmark of 200 dyadic conventions that evaluates compression, acceptance and production through session-isolated evaluation and conjunctive scoring. (2) We provide a diagnostic study across four backbone LLMs and six memory systems that separates memory loss from response behavior. (3) We propose **Pact-Mem**, which preserves partner-specific form–meaning mappings and retrieves them from form-free cues.

2 Related work

Long-term conversation benchmarks. These benchmarks ask for information that some turn of a multi-session history stated. LoCoMo evaluates recall and reasoning over such histories, while LongMemEval and PersonaMem extend the horizon and personalization setting (Maharana et al., 2024; Wu et al., 2025; Jiang et al., 2025). LoCoMo-Plus reduces lexical overlap between a query and its evidence, and T-Mem improves retrieval by writing additional anchors (Li et al., 2026; Guo et al., 2026b). In each case, one turn states the target and fixes its meaning. A dyadic convention instead changes both its form and interpretive evidence across sessions.

Convention formation. Brennan & Clark (1996) define a *conceptual pact*, a partner-specific agreement on how to conceptualize a referent, settled through use and revisable by either party. Repeated interaction contracts referring expressions and makes them specific to the pair (Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996). The in-context conversational adaptation framework (ICCA) evaluates repeated reference games (Hua & Artzi, 2024). Later studies test convention formation across human and LLM dyads and whether their lexical alignment depends on partner history

(Jones et al., 2026; Wang et al., 2026). These games use one session and a public referent whose meaning remains visible to outsiders.

Table 1: Evaluation scope. *Dyad* denotes pair-specific history; stages are probed separately.

	Multi-session	Formation	Dyad	Stage probes
LoCoMo family	✓	×	×	×
Reference games	×	✓	×	×
Supplied meanings (Patel et al., 2023)	×	×	×	×
PactBench (ours)	✓	✓	✓	✓

Closest in semantic structure, MAGNIFICo assigns novel interpretations to expressions, including interpretations presented within long conversations, and tests their downstream use in semantic parsing (Patel et al., 2023). The mapping is supplied rather than collaboratively formed, and the setting is neither partner-specific nor cross-session. Existing work thus tests formation with public referents or tests supplied nonstandard meanings after the mapping is given. PactBench evaluates both formation and later use across sessions (Table 1). Appendix H provides additional comparisons.

3 Preliminaries and Dyadic Conventions

3.1 Preliminaries

Sessions and turns. The interaction history $\mathcal{H}$ of one user and one agent, a *dyad*, contains n timestamped *sessions*. Session s_i, for $1 \le i \le n$, contains m_i *turns*. Turn $u_{i,j}$, for $1 \le j \le m_i$, pairs a speaker with a message:

$$\mathcal{H} = (s_1, \ldots, s_n), \qquad s_i = (u_{i,1}, \ldots, u_{i,m_i}), \qquad u_{i,j} = (a_{i,j},\, x_{i,j}), \quad a_{i,j} \in \{\mathsf{user}, \mathsf{agent}\}.$$

Session s_i carries timestamp τ_i with $\tau_1 < \cdots < \tau_n$, and $\mathcal{H}_{<t}$ and $\mathcal{H}_{\le t}$ are the prefixes before and through s_t. Speakers need not alternate.

Form. A *form* F is a discrete quotable expression used by the dyad to evoke a referent, event, situation, or interpretation, and $|F|$ counts its words. The same referent can be expressed with forms of very different length, for example "the café on Third Street" and "the café".

Meaning. Let $\mathcal{M}(F \mid \mathcal{H})$ be the set of meanings a competent reader recovers for F from history $\mathcal{H}$. The *public-meaning set* $I(F) = \mathcal{M}(F \mid \emptyset)$ contains the meanings recoverable from F without dyad-specific history. Repeated use can establish a *private meaning* $M \in \mathcal{M}(F \mid \mathcal{H}) \setminus I(F)$.

3.2 Dyadic conventions and formation

Dyadic conventions. Following Brennan & Clark (1996), we define a *dyadic convention* as a partner-specific agreement on how to conceptualize a referent. The pair settles it through use rather than by stating it, and its meaning requires their interaction history rather than the form alone. We write a convention as $c = (F, M)$, where M is that private meaning and F is its form:

$$c = (F, M) \quad \text{with} \quad M \in \mathcal{M}(F \mid \mathcal{H}) \setminus I(F).$$

For example, the public meaning of "the bakery" is a shop that sells bread, but within one dyad the same form privately refers to the parking lot behind that shop.

The six-session formation. Prior work describes compression, partner confirmation, and convergence through repeated use (Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996). For controlled evaluation, we instantiate these functions as a *formation history* of six authored reference sessions S_1 to S_6, with formation prefixes $\mathcal{H}_{\le S_1} \subsetneq \cdots \subsetneq \mathcal{H}_{\le S_6}$: $\mathbf{S}_1$ **Origin**, where the referent is described in full; $\mathbf{S}_2$ **Restatement**, where the same description is given in other words; $\mathbf{S}_3$ **Compression**, the first shortened use; $\mathbf{S}_4$ **Compression to bare use**, where the expression arrives with no description attached; $\mathbf{S}_5$ **Transfer**, a bare use on a new occasion; and $\mathbf{S}_6$ **Production**, where the speaker supplies the omitted expression from memory. We use F_k for the form at S_k and $F = F_6$ for the established form. Section 4.1 specifies the user and agent roles that realize this trajectory in the two compression settings.

Conventions vary in how the form and meaning change during formation. Inspired by work on semantic change and repeated reference, we operationalize four recurring mechanisms of dyadic-convention formation (Traugott & Dasher, 2002; Clark & Wilkes-Gibbs, 1986). 1) **Inversion** keeps

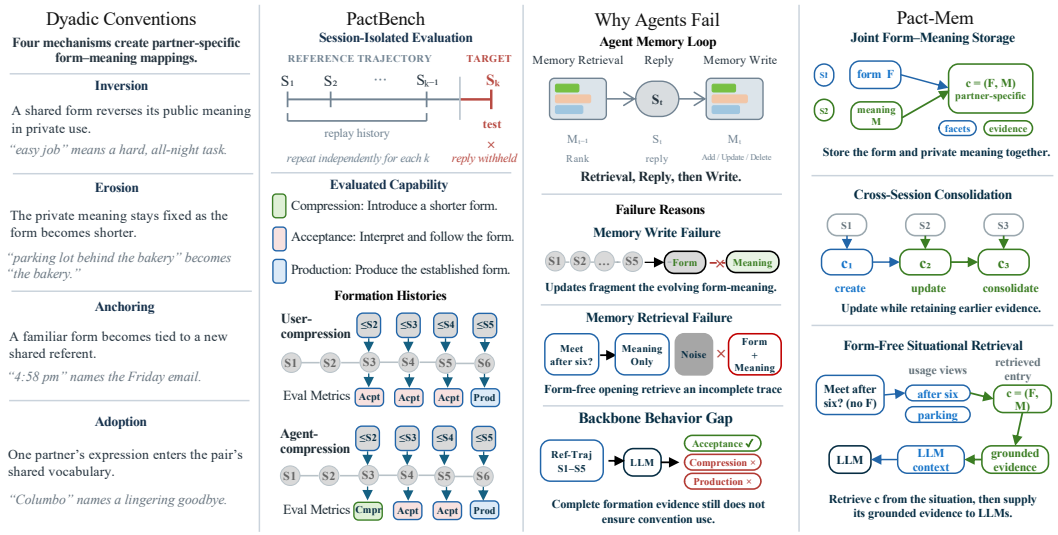

Figure 2: **Overview of PactBench and Pact-Mem.** Dyadic conventions arise through four formation mechanisms (left). PactBench isolates compression, acceptance and production with session-isolated evaluation over user- and agent-compression histories (center-left). The evaluation identifies failures in memory retrieval and agent behavior even when the exact preceding reference formation history is provided (center-right). Pact-Mem combines joint form–meaning storage, cross-session consolidation and form-free situational retrieval (right).

Table 2: Four formation mechanisms over form trajectory $(F_k)_{k=1}^{6}$ and private meaning M.

$\varphi(c)$	Condition	Example
Inversion	F_k fixed, M reverses the polarity of an interpretation in $I(F)$	"career opportunity" $\rightarrow$ a thankless task
Anchoring	F_k fixed, $\exists t : M \notin \mathcal{M}(F \mid \mathcal{H}_{<S_t}),\ M \in \mathcal{M}(F \mid \mathcal{H}_{\leq S_t})$	"Sunny" $\rightarrow$ the friend who cancels
Adoption	F_k fixed, $I(F) \neq \emptyset$, M narrows or extends an interpretation in $I(F)$	"on the road again" $\rightarrow$ the drive back from mom's
Erosion	M fixed, $\lvert F_k \rvert$ decreases until $M \notin I(F_k)$	"almost done?" $\rightarrow$ waiting on the garage-door repair

the form fixed and gives M the opposite polarity of a public interpretation in $I(F)$; the convention forms only because the origin session carried that polarity. 2) **Anchoring** binds M to an unrelated F in one session and never restates it afterwards. 3) **Adoption** takes an expression that is already public and narrows or extends one of its meanings. 4) **Erosion** keeps M fixed and shortens F_k across sessions until the remainder no longer carries M to an outsider. We denote the mechanism of c by $\varphi(c)$. Figure 2 summarizes the benchmark and Pact-Mem, Table 2 compares the four formation conditions, and Appendix B describes their reference trajectories.

4 Evaluation

Given the formation history of Section 3, we evaluate three capabilities the formation calls for: compression (Cmpr), shortening a reference; acceptance (Acpt), confirming the partner's compressed form; and production (Prod), producing the established form when the situation calls for it.

Session-isolated evaluation. An early failure can mask a distinct capability gap at a later formation session. To check which capabilities an agent lacks, we test the capability required at each session independently. At S_k, we follow the reference formation trajectory through S_{k-1}, withhold the reference agent reply r_k, and ask the agent to answer the current user message. The memory system determines the context visible to the backbone LLM when it answers this message.

4.1 Compression settings and session capabilities

Real formation interleaves compression and confirmation between the two speakers. We isolate the agent's roles in two settings: in the *user-compression setting*, the user compresses and the agent accepts; in the *agent-compression setting*, the agent compresses and the user confirms through reuse. Both settings share the first two sessions, which describe the referent in full.

User-compression setting. The user compresses the reference across S_3–S_5: S_3 pairs the form with a partial description of M, S_4 drops the description, and S_5 transfers the bare form to a new situation. The agent accepts each form and produces the established form at S_6.

Agent-compression setting. Compression is not required in every session and is not always the only valid reply. At S_3, a practical contrast requires an unambiguous reference to one of two previously discussed targets without instructing the agent to shorten (Clark & Wilkes-Gibbs, 1986). The reference reply establishes the agent-proposed form; the user reuses it at S_4 and S_5, where the agent must preserve the mapping, and the agent produces it at S_6.

4.2 Conjunctive scoring

Open-ended replies can appear suitable without demonstrating the target capability. We therefore separate semantic judgments from mechanical matching. LLM judges score the *Meaning*, *Response*, and agent-compression S_3 *Compression* checks. Deterministic code scores only the S_6 *Form* check. The Meaning probe asks what the current form F_k means at session k. A dedicated LLM judge tests whether the answer recovers private information supported by the preceding interaction. At every acceptance and production session, the agent must pass both the Meaning check and the corresponding capability check. Agent-compression S_3 uses a unified mapping-and-compression judge.

Compression (S_3 in the agent-compression setting). The agent proposes a shorter form for the meaning established in S_1 and S_2. One answer-key-free LLM judge requires an exact reply span that names the whole target, depends on the private interaction, and leaves at least one established detail implicit. No separate Meaning decision is used because this unified check already tests whether the proposed form preserves the private mapping.

Acceptance (S_3–S_5 in user compression; S_4–S_5 in agent compression). In the user-compression setting, the agent must accept the user's progressively compressed form at S_3, S_4, and S_5. In the agent-compression setting, the agent must accept the user-confirmed form at S_4 and preserve its mapping when the user transfers it to a new situation at S_5. These responses establish the form as common ground (Clark & Schaefer, 1989). An LLM judge implements the Response check, $\mathrm{Resp}(r_k)$, by labeling the reply as *apply*, *generic*, *literal*, or *unrelated*. Apply acts on M; generic remains compatible without applying it; literal uses a public meaning in $I(F)$; and unrelated binds a wrong meaning or answers a different issue. Only apply passes in the user-compression setting because the reply must visibly accept the user's notation. Apply and generic both pass in the agent-compression setting because the user's reuse already confirms the form, while the Meaning check verifies that the agent retains M. Because confirmation occurs at different speakers, the Response criterion differs between the settings. Acceptance scores are therefore comparable within a setting but not directly across settings. The metric requires

$$\mathrm{Acpt}_k = \mathrm{Mean}_k \wedge \mathrm{Resp}(r_k).$$

Production (S_6 in both compression settings). By S_6, both compression settings have established a shared form. The user describes a new situation to which M applies but omits F. Production tests whether the agent retrieves and introduces that form. The same LLM judges provide the Meaning and setting-specific Response decisions. The deterministic Form check, $\mathrm{Form}(r_6)$, requires F, or a registered variant, as a contiguous lowercased word sequence without stemming or paraphrase. This is stricter than a conceptual pact (Brennan & Clark, 1996):

$$\mathrm{Prod}_6 = \mathrm{Mean}_6 \wedge \mathrm{Resp}(r_6) \wedge \mathrm{Form}(r_6).$$

Appendix C reports the Meaning ablation and an audit of 1,600 decisions against a human reference annotator, on which DeepSeek V4 Pro reaches 86.8% agreement.

5 Experiments

5.1 Experimental Setup

Evaluation configuration. The 200 conventions, 50 per formation mechanism, are evaluated in both compression settings and inserted into LoCoMo-Ultra conversations (Appendix B.2 details

the extension and the convention-to-host assignment). The default settings place 120 *pre-sessions* before S_1, insert five host sessions between consecutive formation sessions, and insert no additional *post-sessions* beyond the five host sessions between S_5 and S_6. Model calls use temperature 0 when supported; Kimi K2.6 uses the provider-required non-thinking temperature of 0.6. Appendix D documents the prompts.

Agents and judges. We evaluate Qwen3.8-27B, GPT-4o mini, Kimi K2.6, and DeepSeek V4 Flash as the backbone LLMs, and we build the agent framework on LD-Agent (Li et al., 2025a). DeepSeek V4 Pro judges the semantic checks in Section 4.2.

Memory systems. The memory design determines the context supplied to the agent at each evaluated session. In this paper, we compare the following memory systems: LD-Agent's native memory, BM25 (Robertson & Zaragoza, 2009), dense RAG (Karpukhin et al., 2020), Nano-Memory (Wu et al., 2026c), A-MEM (Xu et al., 2025), Mem0 (Chhikara et al., 2025), and Pact-Mem (Section 6). Dense retrieval uses the all-MiniLM-L6-v2 encoder (Reimers & Gurevych, 2019; Wang et al., 2020). For each LLM-based memory system, the memory LLM is the same as the evaluated agent. Every system updates after every closed session ($N = 1$). Baseline retrievers return the top $k = 5$ sessions at session granularity. Because Pact-Mem entries are shorter, it returns up to 15 entries while remaining below the retrieved-context character budget of five raw sessions (Section 6).

We directly control the prior context to represent three memory-access outcomes. (i) *No-Prior* supplies only the current opening, mimicking a memory system that retrieves nothing. (ii) *Full-History* supplies the unfiltered conversation history without a memory module. (iii) *Ref-Traj* supplies only reference formation sessions $S_1, \ldots, S_{k-1}$ at session S_k, approximating idealized retrieval without unrelated conversation. Appendix E.1 provides a concrete case study.

5.2 Where Existing Memory Systems Fail

Table 3 macro-averages all four backbone LLMs. Across existing memory systems, user compression averages 17.6%, 9.7%, and 11.0% Acpt from S_3 to S_5 and 0.5% $\mathrm{Prod@}S_6$. Agent compression averages 15.8% $\mathrm{Cmpr@}S_3$, 16.1% and 17.6% Acpt at S_4 and S_5, and 0.4% $\mathrm{Prod@}S_6$.

Table 3: Macro-averaged pass rates (%) across four backbone LLMs. Bold marks the best memory system in each column.

User-compression setting				
Condition	$\mathrm{Acpt@}S_3$	$\mathrm{Acpt@}S_4$	$\mathrm{Acpt@}S_5$	$\mathrm{Prod@}S_6$
Full-History	25.4	29.6	24.4	6.1
BM25	**34.1**	**24.6**	**20.8**	**1.6**
Dense RAG	20.9	8.8	13.0	0.0
Nano-Mem	11.1	5.1	9.4	0.0
A-MEM	19.0	7.4	11.0	0.3
Mem0	4.9	2.3	3.8	0.1
LD-Agent	15.5	10.0	8.0	1.3
Agent-compression setting				
Condition	$\mathrm{Cmpr@}S_3$	$\mathrm{Acpt@}S_4$	$\mathrm{Acpt@}S_5$	$\mathrm{Prod@}S_6$
Full-History	13.5	48.4	49.4	7.1
BM25	14.4	**41.4**	**42.6**	**1.3**
Dense RAG	15.1	6.8	7.8	0.1
Nano-Mem	16.0	10.1	13.3	0.0
A-MEM	13.1	18.5	21.6	0.4
Mem0	16.3	5.6	7.1	0.5
LD-Agent	**20.2**	14.4	13.0	0.4

Observation 1: Existing memory systems do not turn repeated exposure into reliable convention use: acceptance remains weak as the form compresses, agent-initiated compression remains limited, and production is nearly absent.

BM25, which uses lexical retrieval over stored raw sessions, leads all five evaluated Acpt points with scores of 20.8%–42.6%. In this benchmark, repeated surface cues support lexical matching, while raw-session storage avoids transformation loss. Dense RAG retrieves the same raw sessions by embedding similarity but reaches only 6.8%–20.9% Acpt. Nano-Mem, A-MEM, Mem0, and LD-Agent instead transform the history through LLM-based filtering, updated notes, extracted facts, or session summaries. Their 20 Acpt cells range from 2.3% to 21.6%, all below Full-History.

Observation 2: Preserving and retrieving raw sessions is more reliable for acceptance than embedding-based retrieval or LLM-generated notes, facts, and summaries.

Table 5: Capability pass rates (%) across backbones under No-Prior, Full-History, and Ref-Traj. Ref-Reply scores the withheld reference replies and passes every check by construction. Bold marks the best generated context for each backbone–capability pair.

		User-compression setting				Agent-compression setting			
Response source	Context	$\mathrm{Acpt@}S_3$	$\mathrm{Acpt@}S_4$	$\mathrm{Acpt@}S_5$	$\mathrm{Prod@}S_6$	$\mathrm{Cmpr@}S_3$	$\mathrm{Acpt@}S_4$	$\mathrm{Acpt@}S_5$	$\mathrm{Prod@}S_6$
Ref-Reply	N/A	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen3.8-27B	No-Prior	0.0	0.0	0.0	0.0	8.0	0.0	0.0	0.0
	Full-History	22.0	26.0	25.0	6.0	16.0	59.0	56.0	6.0
	Ref-Traj	**68.0**	**67.0**	**58.0**	**16.0**	**40.0**	**79.0**	**77.0**	**26.0**
GPT-4o mini	No-Prior	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
	Full-History	8.0	10.0	4.0	0.0	2.0	13.0	10.0	0.0
	Ref-Traj	**39.0**	**45.0**	**38.0**	**15.0**	**8.0**	**55.0**	**60.0**	**14.0**
Kimi K2.6	No-Prior	0.0	0.0	0.0	0.0	14.0	0.0	0.0	0.0
	Full-History	42.0	48.0	41.0	14.0	27.0	68.0	70.0	18.0
	Ref-Traj	**83.0**	**78.0**	**72.0**	**28.0**	**34.0**	**84.0**	**82.0**	**44.0**
DS V4 Flash	No-Prior	0.0	0.0	0.0	0.0	9.5	0.0	0.0	0.0
	Full-History	29.5	34.5	27.5	4.5	9.0	53.5	61.5	4.5
	Ref-Traj	**77.5**	**78.5**	**74.5**	**21.0**	**33.5**	**88.0**	**90.0**	**38.0**

To diagnose why memory fails, Table 4 reports DeepSeek V4 Flash Recall at S_k (the percentage of preceding formation sessions represented in the top-five retrieval) and Hit@S_6 (the percentage of probes retrieving an M-establishing session, an F-using session, or both).

Table 4: DeepSeek V4 Flash retrieval diagnostics (%) for user compression. Recall is formation-session coverage in the top five; Hit@S_6 indicates retrieval of an M-establishing session, an F-using session, or both.

(a) Recall by memory system

Method	S_3	S_4	S_5	S_6
BM25	**66.8**	**47.0**	**41.0**	11.3
Dense RAG	36.0	20.2	24.5	10.4
Nano-Mem	20.2	11.7	20.0	5.8
A-MEM	28.8	13.3	18.2	8.1
Mem0	18.8	11.2	14.1	10.1
LD-Agent	51.0	43.0	35.5	**15.7**

(b) Hit@S_6

Method	M	F	M & F
BM25	14.0	36.5	4.0
Dense RAG	9.0	33.5	2.5
Nano-Mem	15.0	11.0	1.5
A-MEM	5.5	31.5	2.0
Mem0	15.5	29.5	5.0
LD-Agent	**17.5**	**48.5**	**9.0**

BM25 exposes a query-time retrieval failure: it preserves raw sessions and leads recall while the evolving form remains visible. At S_6, however, the opening expresses M without F, and BM25 recall falls to 11.3%. This decline occurs across all four formation mechanisms, although they transform form and meaning in different ways (Appendix E.4). Mem0 and LD-Agent additionally lose information during consolidation: they retain a registered detail of M in 96.5% and 98.5% of stores at S_6, but retain F in only 65.0% and 61.0%. Retrieval creates a broader bottleneck. Under user compression, the four LLM-based memories retrieve both an M-establishing and an F-using session in only 1.5%–9.0% of cases. Under agent compression, repeated uses of the agent-proposed form improve intermediate BM25 recall, but the same pattern holds at the form-free S_6 query: joint provenance hits remain 2.5%–11.0% across memory systems (Appendix E.3). Because provenance identifies the source session rather than its preserved content, both ranges upper-bound usable evidence coverage.

Failure Mode 1: Multi-session formation distributes the private meaning and compressed form across different sessions. Existing memory systems rarely retrieve both by the production session in either compression setting.

5.3 Can Complete Formation Evidence Close the Gap?

Memory systems omit or fragment the evidence supplied to the backbone LLM. Ref-Traj supplies complete formation evidence: the preceding reference formation history without unrelated conversation. Table 5 compares four backbones under No-Prior, Full-History, and Ref-Traj. Ref-Reply reports how the withheld reference replies score under the same protocol.

Across all four backbones, Ref-Traj improves every evaluated capability over Full-History. At S_5, Acpt rises from 4–41% to 38–75% under user compression and from 10–70% to 60–90% under agent compression. Under agent compression, $\mathrm{Cmpr@}S_3$ rises from 2–27% to 8–40%. Production

rises from 0–14% to 15–28% under user compression and from 0–18% to 14–44% under agent compression. However, Ref-Traj remains far below Ref-Reply on agent compression and production.

Observation 3: Complete formation evidence improves all three capabilities but still falls short of Ref-Reply on agent-initiated compression and production.

Failure Mode 2: Because Ref-Traj supplies the complete formation history, its remaining gap to Ref-Reply cannot be explained by storage or retrieval errors. It reflects a backbone-LLM behavior limitation under the tested prompting setup; Pact-Mem addresses the separable memory component.

6 PACT-MEM

Section 5.2 identifies two points at which existing memory systems lose a convention. Session-level updates can separate or overwrite its form and private meaning, while a later opening can express that meaning without repeating the form. Pact-Mem addresses consolidation failure through joint form–meaning storage and cross-session consolidation, and retrieval failure through form-free situational retrieval.

6.1 DESIGN

Joint form–meaning storage. Pact-Mem treats a convention, rather than a session or an isolated fact, as the memory unit. It stores each convention $c = (F, M)$ as one entry, together with its usage conditions and grounded evidence from the dialogue. This convention-centered representation keeps the exact form and its private meaning in the same retrievable unit.

Cross-session evidence consolidation. Because a convention spans sessions, Pact-Mem revises its entry when the form recurs, refining meaning and usage conditions while retaining the expression and prior evidence. A recovery pass proposes missed phrases; the memory model infers their private meanings from closed history without annotations or reference replies.

Form-free situational retrieval. Pact-Mem associates a convention with form-free situational views $V(c)$, enabling retrieval when an opening q describes its usage context but omits F:

$$\rho(c, q) = \max_{v \in V(c)} \cos(E(q), E(v)) .$$

Retrieved entries return the binding and grounded evidence. The views remain hidden from the backbone LLM, and the context gives no instruction to use F; Appendix F details implementation and causal safeguards.

6.2 RECOVERING COMPRESSION, ACCEPTANCE, AND PRODUCTION

Table 6 compares Pact-Mem with BM25, the strongest raw-message retrieval baseline (which preserves raw sessions), and Ref-Traj across four backbone LLMs; Appendix Tables 14 and 15 report all memory systems. Pact-Mem's clearest gain is production. Across all eight backbone–setting pairs, Pact-Mem improves production over BM25 by 3.0–32.0 percentage points. Under Qwen3.8-27B agent compression, $\mathrm{Prod@}S_6$ rises from 1.0% for the strongest existing memory system to 32.0%. The smaller gains for GPT-4o mini are consistent with a coupled-capacity bottleneck: the same LLM serves as both the memory model and the backbone, so limited capacity affects extraction, consolidation, and response generation. Pact-Mem improves every capability on DeepSeek V4 Flash, while its acceptance gains remain mixed across the other backbones.

Table 7 traces the form–meaning binding from stored memory to retrieved context and final production on DeepSeek V4 Flash. Under user compression, BM25, Dense RAG, Nano-Mem, and A-MEM retain raw session text, so their 99.0% *Store both* rate is expected: this check asks whether both elements exist anywhere in memory, not whether retrieval selects them. They retrieve both elements in only 5.5%–38.5% of cases, localizing their evidence-selection loss to retrieval rather than storage. Pact-Mem stores both in 83.5% of cases and retrieves both in 75.5%; under agent compression, it retrieves both in 66.0%, compared with 31.5% for BM25. Even after both elements reach the context, Pact-Mem produces the convention in only 17.2% and 40.2% of eligible user- and agent-compression cases, showing that retrieval does not guarantee response realization.

Table 6: Pact-Mem vs. BM25 and Ref-Traj across backbones. Bold marks the better of BM25 and Pact-Mem; Ref-Traj supplies exact formation evidence.

		User-compression setting				Agent-compression setting			
Backbone	Memory	Acpt@S_3	Acpt@S_4	Acpt@S_5	Prod@S_6	Cmpr@S_3	Acpt@S_4	Acpt@S_5	Prod@S_6
Qwen3.8-27B	BM25	**32.0**	20.0	21.0	2.0	**18.0**	**43.0**	**46.0**	0.0
	Pact-Mem	**32.0**	**33.0**	**25.0**	**14.0**	**18.0**	**43.0**	35.0	**32.0**
	Ref-Traj	68.0	67.0	58.0	16.0	40.0	79.0	77.0	26.0
GPT-4o mini	BM25	**14.0**	**12.0**	**5.0**	0.0	2.0	**21.0**	**23.0**	0.0
	Pact-Mem	12.0	8.0	4.0	**3.0**	**12.0**	**21.0**	22.0	**8.0**
	Ref-Traj	39.0	45.0	38.0	15.0	8.0	55.0	60.0	14.0
Kimi K2.6	BM25	**49.0**	33.0	30.0	4.0	24.0	**52.0**	**52.0**	4.0
	Pact-Mem	**49.0**	**42.0**	**36.0**	**12.0**	**62.0**	41.0	40.0	**18.0**
	Ref-Traj	83.0	78.0	72.0	28.0	34.0	84.0	82.0	44.0
DS V4 Flash	BM25	41.5	33.5	27.0	0.5	13.5	49.5	49.5	1.0
	Pact-Mem	**53.0**	**55.5**	**37.5**	**13.0**	**48.0**	**54.0**	**54.5**	**26.5**
	Ref-Traj	77.5	78.5	74.5	21.0	33.5	88.0	90.0	38.0

Table 7: Content diagnostics at S_6 on DeepSeek V4 Flash (%). Store/Retrieve both denote the exact form and private meaning in memory/context; conditional production is successes/eligible cases. Bold marks maximum coverage. Appendix Table 19 reports marginal recall.

(a) User-compression setting

Method	Store both	Retrieve both	Prod given both retrieved
BM25	**99.0**	38.5	1.3 (1/77)
Dense RAG	**99.0**	29.0	0.0 (0/58)
Nano-Mem	**99.0**	5.5	0.0 (0/11)
A-MEM	**99.0**	26.5	0.0 (0/53)
Mem0	64.5	3.0	16.7 (1/6)
LD-Agent	60.0	9.0	11.1 (2/18)
Pact-Mem	83.5	**75.5**	17.2 (26/151)

(b) Agent-compression setting

Method	Retrieve both	Prod given both retrieved
BM25	31.5	1.6 (1/63)
Dense RAG	5.5	9.1 (1/11)
Nano-Mem	4.5	0.0 (0/9)
A-MEM	31.0	1.6 (1/62)
Mem0	5.0	0.0 (0/10)
LD-Agent	15.5	3.2 (1/31)
Pact-Mem	**66.0**	40.2 (53/132)

Ref-Traj remains stronger on acceptance for every backbone and setting and on production in seven of eight pairs. Pact-Mem exceeds Ref-Traj on Qwen3.8-27B agent-compression production and on compression for three backbones, showing that structured retrieval can alter responses even when Ref-Traj supplies all formation sessions. The remaining gaps are consistent with the backbone-LLM limitation in Section 5.3.

6.3 Post-Formation Retention

Long-term memory should retain established conventions across later interaction. Holding the main configuration fixed, we insert 40, 80, or 120 host sessions before S_6 and compare LD-Agent, BM25, and Mem0, the baselines with nonzero zero-delay production. Other baselines, including A-MEM, already score near zero in Table 3, leaving little production to retain.

Table 8: Post-formation production (Prod@S_6, %) on 200 DeepSeek V4 Flash user-compression trajectories. Baselines are the three strongest at zero delay.

Memory	Additional sessions			
	0	40	80	120
LD-Agent	1.0	1.5	0.0	0.0
BM25	0.5	0.5	0.5	0.5
Mem0	0.5	0.0	0.0	0.0
Pact-Mem	**13.0**	**13.5**	**11.5**	**10.5**

Pact-Mem retains 13.5% production after 40 sessions and 10.5% after 120, compared with 13.0% at zero delay (Table 8). BM25 remains at 0.5%; Mem0 and LD-Agent reach 0.0% by 40 and 80 sessions, respectively. This modest decline does not distinguish storage from retrieval loss. Pact-Mem is orthogonal to expiration, selective forgetting, and other memory-update modules, which can coexist with its convention-level updates to control bounded growth.

7 Conclusion

PactBench evaluates whether long-term agents form and use dyad-specific conventions. Across four LLM backbones, existing memories separate form from private meaning, while complete formation evidence shows a distinct response-behavior limitation. Pact-Mem targets the memory failure through joint form–meaning storage, cross-session consolidation and form-free retrieval, improving production by up to 31 percentage points.

AI USE STATEMENT

For manuscript preparation, we used LLMs for paper polishing only. For data construction, Claude Sonnet 5 generated candidate conventions and expanded formation histories in both compression settings; Qwen3.8-27B, DeepSeek V4 Flash, GPT-4o mini, and Kimi K2.6 served as the evaluated backbone LLMs and as the memory models inside LLM-based memory systems. DeepSeek V4 Pro judged the semantic checks.

REPRODUCIBILITY STATEMENT

Section 4 defines the evaluation and scoring rules, and Section 5.1 specifies the timeline, models, cohort sizes and run settings. Appendix B documents benchmark construction, while Appendices C and D report judge agreement and prompt routing. Code and data will be made available upon acceptance.

REFERENCES

Hirona Jacqueline Arai and Xiang Ren. Drinq: Evaluating conversational implicature with controlled context variation, 2026. URL https://arxiv.org/abs/2605.24267.

Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996. doi: 10.1037/0278-7393.22.6.1482.

Yuxuan Cai, Wei Li, Jie Zhou, Qin Chen, Xin Li, Bo Zhang, and Liang He. Ask only when needed: Proactive retrieval from memory and skills for experience-driven lifelong agents, 2026. URL https://arxiv.org/abs/2604.20572.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13(2): 259–294, 1989. doi: 10.1207/s15516709cog1302_7.

Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1): 1–39, 1986. doi: 10.1016/0010-0277(86)90010-7.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. WinoDict: Probing language models for in-context word acquisition. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 94–102, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.7. URL https://aclanthology.org/2023.eacl-main.7/.

Zhaoxin Feng, Jianfei Ma, and Emmanuele Chersoni. Know it, act on it: Investigating memory utilization in llm personalization, 2026. URL https://arxiv.org/abs/2607.29433.

Xing Fu, Yulin Hu, Mengtong Ji, Haozhen Li, Yixin Sun, Weixiang Zhao, Yanyan Zhao, and Bing Qin. Enpmr-bench: Benchmarking proactive memory retrieval for emotional support agents, 2026. URL https://arxiv.org/abs/2605.27240.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. Towards natural personalization: Evaluating long-horizon preference following in personalized user-llm interactions, 2026a. URL https://arxiv.org/abs/2603.04191.

Weidong Guo, Dakai Wang, Zixuan Wang, Hui Liu, and Yu Xu. T-mem: Memory that anticipates, not archives, 2026b. URL https://arxiv.org/abs/2606.15405.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL https://aclanthology.org/2023.acl-long.230/.

Yilun Hua and Yoav Artzi. Talk less, interact better: Evaluating in-context conversational adaptation in multimodal llms, 2024. URL https://arxiv.org/abs/2408.01417.

Yilun Hua, Evan Wang, and Yoav Artzi. Post-training for efficient communication via convention formation, 2026. URL https://arxiv.org/abs/2508.06482.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory, 2025. URL https://arxiv.org/abs/2512.06688.

Cameron R. Jones, Agnese Lombardi, Kyle Mahowald, and Benjamin K. Bergen. Llms and people both learn to form conventions – just not with each other, 2026. URL https://arxiv.org/abs/2602.08208.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A. Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. Acon: Optimizing context compression for long-horizon llm agents, 2026. URL https://arxiv.org/abs/2510.00615.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020. URL https://arxiv.org/abs/2004.04906.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! LLM-powered personalized agent for long-term dialogue. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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. 5259–5276, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.272. URL https://aclanthology.org/2025.naacl-long.272/.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents, 2026. URL https://arxiv.org/abs/2602.10715.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhen Tao, Huayi Lai, Hao Wu, Bo Tang, Zhengren Wang, Zhaoxin Fan, Ningyu Zhang, Linfeng Zhang, Junchi Yan, Mingchuan Yang, Tong Xu, Wei Xu, Huajun Chen, Haofen Wang, Hongkang Yang, Wentao Zhang, Zhi-Qin John Xu, Siheng Chen, and Feiyu Xiong. Memos: A memory os for ai system, 2025b. URL https://arxiv.org/abs/2507.03724.

Genglin Liu and Saadia Gabriel. Pm-bench: Evaluating prospective memory in llm agents, 2026. URL https://arxiv.org/abs/2607.12385.

Yao Liu, Guangjia Chai, Yuming Huang, Jihao Huang, Lei Wang, and Junchen Wan. Companionbench: A theory-anchored, real-world-grounded benchmark for ai emotional companionship, 2026. URL https://arxiv.org/abs/2608.02046.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents, 2024. URL https://arxiv.org/abs/2402.17753.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. On memory construction and retrieval for personalized conversational agents, 2025. URL https://arxiv.org/abs/2502.05589.

Arkil Patel, Satwik Bhattamishra, Siva Reddy, and Dzmitry Bahdanau. MAGNIFICo: Evaluating the in-context learning ability of large language models to generalize to novel interpretations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2167–2189, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.134. URL https://aclanthology.org/2023.emnlp-main.134/.

Christian Poelitz, Finale Doshi-Velez, and Siân Lindley. A benchmark to assess common ground in human-ai collaboration, 2026. URL https://arxiv.org/abs/2602.21337.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Stephen E. Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. Scaling long-horizon llm agent via context-folding, 2025. URL https://arxiv.org/abs/2510.11967.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.989. URL https://aclanthology.org/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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

Elizabeth Closs Traugott and Richard B. Dasher. *Regularity in Semantic Change*. Cambridge University Press, Cambridge, 2002. doi: 10.1017/CBO9780511486500.

Saujas Vaduguru, Yilun Hua, Yoav Artzi, and Daniel Fried. Success and cost elicit convention formation for efficient communication, 2025. URL https://arxiv.org/abs/2510.24023.

Patricio M. Vera. Lexical consensus: Grounded word learning and shared meaning in artificial agents, 2026. URL https://arxiv.org/abs/2606.22207.

Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu,

Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Po-Ya Angela Wang, Chinmaya Mishra, Aslı Özyürek, Paula Rubio-Fernández, and Esam Ghaleb. Aligned but not partner-specific: Distinguishing how multimodal llm agents succeed in reference games without human-like conventions, 2026. URL https://arxiv.org/abs/2606.08081.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents, 2025. URL https://arxiv.org/abs/2507.07957.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory, 2025. URL https://arxiv.org/abs/2410.10813.

Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang, and Ronghao Chen. KnowMe-bench: Benchmarking person understanding for lifelong digital companions. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30214–30238, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1394. URL https://aclanthology.org/2026.acl-long.1394/.

Yifan Wu, Lizhu Zhang, Yuhang Zhou, Mingyi Wang, Bo Peng, Serena Li, Xiangjun Fan, and Zhuokai Zhao. Remember when it matters: Proactive memory agent for long-horizon agents, 2026b. URL https://arxiv.org/abs/2607.08716.

Yuqian Wu, Wei Chen, Zhengjun Huang, Junle Chen, Qingxiang Liu, Kai Wang, Xiaofang Zhou, and Yuxuan Liang. Back to basics: Let conversational agents remember with just retrieval and generation, 2026c. URL https://arxiv.org/abs/2604.11628.

Yuqian Wu, Zhijie Deng, Wei Chen, Junwei Li, Yutian Jiang, Junle Chen, Zhengjun Huang, Qingxiang Liu, Jing Tang, Jiaheng Wei, and Yuxuan Liang. Lifeside: Benchmarking agents as lifelong digital companions, 2026d. URL https://arxiv.org/abs/2606.04660.

Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Alpsbench: An llm personalization benchmark for real-dialogue memorization and preference alignment, 2026. URL https://arxiv.org/abs/2603.26680.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey, 2024. URL https://arxiv.org/abs/2403.08319.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025. URL https://arxiv.org/abs/2502.12110.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in llms, 2024. URL https://arxiv.org/abs/2405.08760.

Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models, 2026. URL https://arxiv.org/abs/2505.18497.

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo, and Yafeng Deng. Hypermem: Hypergraph memory for long-term conversations, 2026. URL https://arxiv.org/abs/2604.08256.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023. URL https://arxiv.org/abs/2305.10250.

APPENDIX

A LIMITATIONS

We acknowledge the following limitations. (1) PactBench uses synthetic conventions and controlled formation trajectories rather than conventions collected from long-term human–agent conversations, as collecting controlled interactions over a two-year horizon is impractical. (2) Pact-Mem focuses on the memory component. Addressing the remaining response-behavior gap may require backbone-model training, which is complementary to this work.

B BENCHMARK CONSTRUCTION

B.1 CONVENTION CARDS AND FORMATION HISTORIES

Pool. PactBench holds 200 conventions, 50 per formation mechanism, each carrying the two formation histories of Section 3.2, so the pool realizes 400 formation histories. A convention is keyed by (F, M), so one form carrying two private meanings gives two conventions, judged independently. Candidates are generated one mechanism at a time by a Claude Sonnet 5 generator (Introducing Claude Sonnet 5) and screened by the gates below; every rejected candidate is kept with its reason. The conventions used to calibrate the response judge are held out from every reported score. Each convention is generated and screened against one mechanism-specific formation condition; the released data assigns exactly one mechanism label to each convention. Because erosion shortens F across sessions, we score each formation session separately. We report scores both across the full pool and separately by mechanism. Mechanism constrains what an outsider recovers, and the pool composition is: inversion reverse 30, wrong 20; anchoring nothing 24, wrong 26; adoption wrong 26, coarser 24; erosion nothing 25, wrong 19, coarser 6.

Ingestion gates. The public-meaning set $I(F)$ used by the gates below is obtained by sampling a pinned strong model on the form alone and applying an independent public-reading screen. A candidate is admitted only if it passes seven gates. Four govern what the material may give away: (i) no probe solvable by time anchoring, local context, or public common sense; (ii) $M \notin I(F)$; (iii) a turn may tell the origin event verbatim, but no turn may state the pair (F, M), so M is inducible only across turns and never retrievable as a sentence; (iv) F is quotable in the sense of Section 3.1, which puts operators over arbitrary content out of scope. The remaining gates require M to differ from every public interpretation in propositional content or speech act and exclude mappings recoverable from temporal, absence, sequence, or other superficial cues. The timeline in Section 5.1 sets the interval between sessions.

Leakage audit. Leakage is audited at three points. First, the gates inspect the origin narrative and the form. Second, every generated session is scanned for turns that pair F with a gloss of M; any such turn is regenerated. Third, a reader with no access to the pair must fail to recover M from a probe alone. The No-Prior condition of Section 5.1 is the empirical counterpart of these checks.

Formation scripts. Each formation history is a script of six sessions realizing the trajectory of Section 3.2 with a fixed shape, 5 to 6 turns at $k = 1$ shrinking to 1 to 2 at $k = 6$. Automated gates

check the turn counts, the placement of F, glossing and elapsed-time phrases, and an exemplar self-check verifies that an ideal hand-written convention passes every gate and that the reference turns pass the metric layers of Section 4.2.

Probe authoring. The production probe must fit M without stretching it, differ from the origin event, and not disclose M to a stranger. It must read like an ordinary conversational turn and carry no task label. In the running example, the probe asks where to leave the car. A probe whose opening already contains the form is unscorable, so weaving anomalies cannot create credit.

B.2 LoCoMo-Ultra host corpus

LoCoMo-Ultra preserves all 262 sessions from the ten LoCoMo dyads (Maharana et al., 2024). It adds 3,261 generated sessions, producing 3,523 sessions and 69,900 turns. Each host contains 348–357 sessions, spans 724–734 days, and contains 218K–245K rendered tokens. The original sessions remain verbatim.

The extension follows four stages. Seed extracts a typed profile; grow creates dated events with explicit state updates; expand realizes each accepted event as one chat session without adding profile facts; and audit checks speaker identity, chronology, slot consistency, and references to ended items. Code rejects illegal updates before generation, while timestamps and assembly are deterministic. The typed slots distinguish mutable values, immutable lifetime values, sets, and counters. These checks cover the slot inventory rather than every detail in the generated dialogue.

B.3 Convention assignment and insertion

Each convention is assigned to one host before evaluation, with 20 conventions per host. Each convention then defines an independent evaluation timeline copied from its assigned host; the 20 conventions are not inserted into one shared evaluation history. The frozen assignment minimizes exact occurrences of the form under this load constraint. Among 200 conventions, 193 have no exact-form occurrence in their assigned host, and the maximum collision count is three. Each result row records this count as an interference covariate.

Each formation exchange continues an existing host session and inherits its timestamp. In the frozen main configuration, S_1 follows 120 host sessions and five host sessions lie between consecutive formation sessions. The user-compression and agent-compression versions of one convention use the same frozen host assignment.

B.4 Fairness protocol for memory modules

Every store runs on the same timeline, the same memory operation interval, and the same agent shell, with equal and reported consolidation counts. Each probe runs before the store observes the probe session, and no store may branch on which sessions are formation sessions. Every result row records through which channel the form reached the backbone LLM: the store's retrieved units, the buffer of sessions not yet consolidated, or the context prefix.

C Evaluation validity

C.1 Effect of the Meaning check

The Meaning check prevents a locally plausible reply from receiving credit when the agent does not recover the private mapping. Table 9 separates three quantities for the DeepSeek V4 Flash runs: the semantic Meaning check (Sem.), the capability check without Meaning (Cap.), and their reported conjunction (Joint). For acceptance, Cap. is the Response check. For production, Cap. is Response $\wedge$ Form. Agent-compression S_3 uses one unified Compression judge and therefore has only Joint.

Table 9: Effect of the Meaning check on DeepSeek V4 Flash pass rates (%). Sem. is Meaning, Cap. is the corresponding capability check without Meaning, and Joint is the reported joint score. Agent-compression S_3 is not decomposable because it uses a unified judge. Bold marks the highest generated Joint score at each session. At S_k, Ref-Traj contains only reference sessions $S_1, \ldots, S_{k-1}$ and excludes all LoCoMo-Ultra host sessions.

	S_3			S_4			S_5			S_6		
Condition	Sem.	Cap.	Joint	Sem.	Cap.	Joint	Sem.	Cap.	Joint	Sem.	Cap.	Joint
User-compression setting												
No-Prior	1.5	5.0	0.0	0.5	2.5	0.0	0.0	1.5	0.0	0.0	0.0	0.0
Ref-Traj	96.5	81.0	**77.5**	99.0	79.0	**78.5**	99.5	74.5	**74.5**	98.5	21.0	21.0
Full-History	69.0	41.5	29.5	71.5	43.0	34.5	75.0	35.0	27.5	74.0	5.0	4.5
BM25	72.0	46.5	41.5	58.0	41.5	33.5	53.0	34.0	27.0	12.5	0.5	0.5
Dense RAG	50.0	30.0	27.5	24.5	21.5	13.0	41.5	21.5	17.0	9.0	0.0	0.0
Nano-Mem	23.0	21.5	14.5	13.0	16.5	8.5	26.0	18.0	12.5	4.0	0.0	0.0
A-MEM	40.0	27.5	24.0	17.0	14.5	8.5	30.5	19.5	13.0	6.0	0.5	0.0
Mem0	13.0	16.0	6.5	7.0	9.5	2.0	11.0	11.5	6.0	1.0	1.0	0.5
LD-Agent	34.0	32.0	22.0	26.0	20.0	13.0	25.0	17.5	9.0	6.0	2.0	1.0
Pact-Mem	68.5	58.5	**53.0**	78.5	60.0	**55.5**	61.5	47.5	**37.5**	47.5	19.5	**13.0**
Agent-compression setting												
No-Prior	–	–	9.5	0.0	74.5	0.0	0.0	66.0	0.0	0.0	0.0	0.0
Ref-Traj	–	–	33.5	92.5	95.0	**88.0**	95.0	93.5	**90.0**	95.0	39.0	38.0
Full-History	–	–	9.0	66.0	78.5	53.5	70.5	83.5	61.5	75.0	4.5	4.5
BM25	–	–	13.5	55.5	80.5	49.5	56.0	80.5	49.5	19.0	2.0	1.0
Dense RAG	–	–	22.5	7.5	67.5	6.0	11.5	64.5	9.0	2.5	0.5	0.5
Nano-Mem	–	–	23.0	14.0	70.0	12.5	21.5	71.5	18.0	4.0	0.0	0.0
A-MEM	–	–	15.5	28.0	70.0	26.0	36.5	71.5	30.5	17.0	1.5	0.5
Mem0	–	–	22.0	8.0	67.5	7.5	12.0	62.5	9.5	4.0	0.5	0.0
LD-Agent	–	–	25.6	25.5	72.0	23.5	24.5	65.0	19.0	10.5	2.0	0.5
Pact-Mem	–	–	**48.0**	59.0	77.0	**54.0**	60.5	79.5	54.5	43.0	38.0	**26.5**

No-Prior passes the agent-compression Response check on 74.5% and 66.0% of replies at S_4 and S_5, but joint acceptance is 0% because the private mapping is absent. Ref-Traj keeps Meaning above 92% at every decomposable session, showing that the Meaning conjunct removes plausible guesses without causing the low production rates.

C.2 Agreement with a human reference annotator

We audit 1,600 capability decisions: 200 conventions, two compression settings and four evaluated sessions. Within each formation mechanism, 50 cards were randomly split into two 25-card cohorts with opposite BM25 and Pact-Mem assignments across the two settings, which yields 100 cases per setting–session–memory cell. A human reference annotator trained on the benchmark task and frozen scoring rubric labeled the same 1,600 cases. For each case, the annotator received the same task evidence and rubric as the corresponding semantic judge, including the registered target annotation only when that judge received it. Neither received the reference reply. The annotator was also blinded to the memory-condition name, DeepSeek V4 Pro judgment, and automated pipeline decision. Table 10 reports the DeepSeek V4 Pro pass rate and per-case agreement with the human reference labels.

Table 10: Agreement between DeepSeek V4 Pro and a human reference annotator (%). Human and DS are pass rates; Agreement is measured per case. Each row contains 200 cases.

(a) User compression

Capability	Human	DS	Agreement
$\mathrm{Acpt@}S_3$	54.5	43.0	77.5
$\mathrm{Acpt@}S_4$	54.0	42.5	81.5
$\mathrm{Acpt@}S_5$	43.0	34.5	82.0
$\mathrm{Prod@}S_6$	8.5	7.0	**98.0**

(b) Agent compression

Capability	Human	DS	Agreement
$\mathrm{Cmpr@}S_3$	17.0	30.5	84.5
$\mathrm{Acpt@}S_4$	47.5	49.5	90.0
$\mathrm{Acpt@}S_5$	52.0	53.5	83.5
$\mathrm{Prod@}S_6$	13.5	12.0	**97.5**

DeepSeek V4 Pro agrees with the human reference annotator on 1,389 of 1,600 decisions (86.8%). Agreement reaches 97.5%–98.0% for production, 77.5%–90.0% for acceptance and 84.5% for unified Compression. The audit provides an independent human cross-check of the reported benchmark

scores and identifies acceptance decisions as the main remaining source of judge–human disagreement.

D PROMPT SPECIFICATIONS

Table 11 groups every LLM prompt family used in benchmark construction, backbone generation, memory processing and evaluation. Construction prompts receive the registered convention because they author or validate the benchmark. Experimental prompts receive only the context allowed by their condition.

Table 11: Prompt inventory. The information column lists the task-specific content supplied in addition to the fixed system instruction.

Prompt family	Purpose	Task-specific information
Convention creation	Generate and screen a convention card	Formation mechanism, form–meaning constraints, public-reading tests and structured output schema
Formation expansion	Author the two six-session histories	Accepted card, mechanism, speaker roles, session trajectory and assigned LoCoMo-Ultra profile
Backbone reply	Generate the evaluated agent turn	Retrieved memory, recent unconsolidated sessions and current opening; no registered convention or reference reply
Convention-use guidance	Test the response-behavior intervention	The shared backbone prompt plus one condition-independent instruction quoted below
Meaning probe	Recover the private interpretation	The same available context, current form and a matched no-history query
Semantic judges	Score Meaning, Response and Compression	Inputs and structured decisions specified in Table 12
Memory processing	Construct or filter stored context	Closed sessions and each method's published or released extraction schema
Pact-Mem processing	Extract and consolidate convention entries	Closed-session window, joint form–meaning schema and recurring phrases proposed by the recovery pass

D.1 CONSTRUCTION AND BACKBONE PROMPTS

Construction follows the routing in Table 11; the automated gates in Appendix B validate every generated card and formation history.

Backbone generation. The shared system prompt for the backbone LLM gives no instruction to remember, shorten, accept, or produce a convention. Pact-Mem reports the retrieved form–meaning binding and its grounded evidence as factual memory, without instructing the backbone LLM to use the form. Every reply uses the following system message:

> You are texting with your long-term partner. You know each other well and text casually. You are an online chat companion: you exist only inside this chat, you never go anywhere, never handle a physical object, and are never in the same place as your partner. Continue the conversation with your next message. Reply with just the message, nothing else.

The user message concatenates retrieved memory, unconsolidated recent sessions, and the open formation session, then ends with `You:`. No reply-length instruction is added; the 2,048-token output ceiling is not a target length.

The **Pact-Mem + Convention-Use Guidance** condition appends: "Write the way the two of you actually text each other, not the way you would write to someone you had just met. Where the two of you already have a way of referring to something, use it." This condition is excluded from Table 5.

D.2 EVALUATION PROMPTS

Meaning probe. The Meaning probe uses the same available context as the reply probe. Its system message is:

> You are reading your own text conversation with your long-term partner. Answer the question at the end in one or two plain sentences, using what you remember of your chats. If you do not know, say so.

The probe appends: "Question: your partner just wrote the message above. What does `<form>` refer to between the two of you? One or two plain sentences, in your own words." A matched no-history copy receives the current session and the same question. The Meaning judge uses this answer to test whether the full-context answer adds supported private information beyond the no-history answer.

Judge routing. Table 12 records each judge input and output. All judge calls request temperature zero and JSON output; arm names, earlier labels, and mechanical pass bits are withheld.

Table 12: Prompt routing for semantic judgments.

Prompt	Inputs	Required output
User Meaning	Form, question, meaning at the current cut, facets, inversion flag, prior formation sessions, matched no-history answer, evaluated answer	Meaning class, role correctness, inversion polarity, beyond-no-history bit, and exact answer evidence
User Response	Form, private meaning, facet table, prior formation sessions, current session, opening, evaluated reply	Physical-action check, opening facts, private contribution, contradiction, concept keys, reason, and one of apply, generic, literal, or unrelated
Agent S_3	S_1, S_2, the S_3 user message, and the evaluated reply	One exact reply span that passes whole-target reference, private mapping, and compression; otherwise fail
Agent S_4–S_6 Meaning	Current form, target annotation, full available context, matched no-history context, and both answers	Correct target binding, beyond-no-history bit, and evidence IDs
Agent S_4–S_6 Response	Current form, target annotation, available context, and evaluated reply	Apply, generic, literal, or unrelated, with exact reply evidence

The user Response judge passes only *apply*. Apply requires a private detail or implication in the reply that is absent from the current opening. Generic compatibility, the public reading, an unrelated answer, a contradiction, or a physical action attributed to the online companion does not pass. The agent-compression setting also accepts a generic response at S_4–S_6 because the user has already reused the agent's proposed form; its separate Meaning check verifies that the mapping is retained.

The agent S_3 judge is answer-key-free. It must cite one exact expression in the reply that names the whole target, depends on S_1 and S_2 for its mapping, and leaves at least one formation detail implicit. Merely following the convention's advice, naming one component, or using a publicly readable expression fails.

D.3 MEMORY PROMPTS

No-Prior, Full-History, Ref-Traj, BM25, dense RAG, and Nano-TIR add no generative memory prompt. Nano-Memory first retrieves sessions with Nano-TIR, then gives the query and those sessions to one filter call with this system message:

> Below are excerpts from two people's chat history, followed by the message one of them is about to answer. Copy out only the parts of the history that are relevant to that message. Copy them word for word, keep them in the order they appear, and drop everything else: greetings, sign-offs, topic transitions, and exchanges about other matters. Reply with just the extracted fragments and nothing else.

The remaining baselines use their released or default prompts: LD-Agent writes a bounded factual summary, A-MEM writes and revises structured notes, and Mem0 uses its upstream extraction and

update prompts. Every method sees the same permitted conversation state and passes its output to the shared backbone prompt.

Pact-Mem processing. Pact-Mem extracts the joint form–meaning entry defined in Appendix F. Its prompts receive only closed dialogue and never the registered convention, reference reply, or evaluation label.

E EVALUATION CONTEXTS AND FULL RESULTS

E.1 WHAT REF-TRAJ SUPPLIES

At S_k, Ref-Traj supplies $S_1, \ldots, S_{k-1}$, excludes host sessions, appends the shared current opening, and withholds the reference reply. Table 13 illustrates this construction; its registered form–meaning annotation explains the example but is not supplied to the backbone LLM.

Table 13: Ref-Traj at the S_6 production evaluation for one erosion convention. Quoted text abbreviates the complete sessions supplied to the backbone LLM.

Component	Content and role
Convention	"Bell Street?" refers to the half chicken from the charcoal-grill place on Bell Street. This annotation is used for evaluation but is not supplied directly.
S_1–S_2	The partner introduces the half chicken and garlic sauce, then returns for "Bell Street chicken again." These sessions establish the private mapping.
S_3–S_5	The partner progressively shortens the reference to "Bell Street?" and reuses it when hungry and when planning a visit with their sister.
S_6 opening	"Took my sister to the charcoal grill place last night." This current message contains neither the established form nor the complete private meaning.
Withheld reply	The authored agent turn begins "so she finally got the Bell Street experience." The backbone LLM must produce its own reply instead.

No-Prior supplies no earlier sessions, Full-History supplies the unfiltered conversation, and each memory system supplies its retrieved context. Ref-Traj removes memory selection and transformation errors while leaving response generation to the backbone LLM.

E.2 FULL CAPABILITY RESULTS

Tables 14 and 15 report every memory system and capability for all four backbone LLMs. These results identify the strongest existing memory baseline for each backbone and capability rather than inferring it from the focused comparison in Table 6. Table 16 then splits Ref-Traj by formation mechanism for the DeepSeek V4 Flash backbone. DeepSeek V4 Pro judges all conditions.

Table 14: Complete capability pass rates (%) in the user-compression setting. Full-History is a context baseline; Ref-Traj supplies exact formation evidence. Bold marks Pact-Mem where it matches or exceeds the strongest existing memory system.

Backbone	Context	$\mathrm{Acpt@}S_3$	$\mathrm{Acpt@}S_4$	$\mathrm{Acpt@}S_5$	$\mathrm{Prod@}S_6$
Qwen3.8-27B	Full-History	22.0	26.0	25.0	6.0
	BM25	32.0	20.0	21.0	2.0
	Dense RAG	15.0	6.0	10.0	0.0
	Nano-Mem	12.0	3.0	7.0	0.0
	A-MEM	16.0	7.0	6.0	0.0
	Mem0	4.0	1.0	4.0	0.0
	LD-Agent	8.0	7.0	6.0	0.0
	Pact-Mem	**32.0**	**33.0**	**25.0**	**14.0**
	Ref-Traj	68.0	67.0	58.0	16.0
GPT-4o mini	Full-History	8.0	10.0	4.0	0.0
	BM25	14.0	12.0	5.0	0.0
	Dense RAG	8.0	2.0	4.0	0.0
	Nano-Mem	2.0	1.0	1.0	0.0
	A-MEM	6.0	2.0	3.0	0.0
	Mem0	0.0	2.0	0.0	0.0
	LD-Agent	2.0	0.0	0.0	0.0
	Pact-Mem	12.0	8.0	4.0	**3.0**
	Ref-Traj	39.0	45.0	38.0	15.0
Kimi K2.6	Full-History	42.0	48.0	41.0	14.0
	BM25	49.0	33.0	30.0	4.0
	Dense RAG	33.0	14.0	21.0	0.0
	Nano-Mem	16.0	8.0	17.0	0.0
	A-MEM	30.0	12.0	22.0	1.0
	Mem0	9.0	4.0	5.0	0.0
	LD-Agent	30.0	20.0	17.0	4.0
	Pact-Mem	**49.0**	**42.0**	**36.0**	**12.0**
	Ref-Traj	83.0	78.0	72.0	28.0
DS V4 Flash	Full-History	29.5	34.5	27.5	4.5
	BM25	41.5	33.5	27.0	0.5
	Dense RAG	27.5	13.0	17.0	0.0
	Nano-Mem	14.5	8.5	12.5	0.0
	A-MEM	24.0	8.5	13.0	0.0
	Mem0	6.5	2.0	6.0	0.5
	LD-Agent	22.0	13.0	9.0	1.0
	Pact-Mem	**53.0**	**55.5**	**37.5**	**13.0**
	Ref-Traj	77.5	78.5	74.5	21.0

Table 15: Complete capability pass rates (%) in the agent-compression setting. Full-History is a context baseline; Ref-Traj supplies exact formation evidence. Bold marks Pact-Mem where it matches or exceeds the strongest existing memory system.

Backbone	Context	$\mathrm{Cmpr@}S_3$	$\mathrm{Acpt@}S_4$	$\mathrm{Acpt@}S_5$	$\mathrm{Prod@}S_6$
Qwen3.8-27B	Full-History	16.0	59.0	56.0	6.0
	BM25	18.0	43.0	46.0	0.0
	Dense RAG	16.0	8.0	6.0	0.0
	Nano-Mem	19.0	12.0	14.0	0.0
	A-MEM	15.0	20.0	23.0	1.0
	Mem0	19.0	8.0	6.0	0.0
	LD-Agent	30.0	14.0	12.0	0.0
	Pact-Mem	18.0	**43.0**	35.0	**32.0**
	Ref-Traj	40.0	79.0	77.0	26.0
GPT-4o mini	Full-History	2.0	12.0	11.0	0.0
	BM25	2.0	21.0	23.0	0.0
	Dense RAG	2.0	5.0	4.0	0.0
	Nano-Mem	2.0	1.0	5.0	0.0
	A-MEM	3.0	10.0	9.0	0.0
	Mem0	4.0	3.0	2.0	0.0
	LD-Agent	1.0	2.0	4.0	0.0
	Pact-Mem	**12.0**	**21.0**	22.0	**8.0**
	Ref-Traj	8.0	55.0	60.0	14.0
Kimi K2.6	Full-History	27.0	68.0	70.0	18.0
	BM25	24.0	52.0	52.0	4.0
	Dense RAG	20.0	8.0	12.0	0.0
	Nano-Mem	20.0	15.0	16.0	0.0
	A-MEM	19.0	18.0	24.0	0.0
	Mem0	20.0	4.0	11.0	2.0
	LD-Agent	24.0	18.0	17.0	1.0
	Pact-Mem	**62.0**	41.0	40.0	**18.0**
	Ref-Traj	34.0	84.0	82.0	44.0
DS V4 Flash	Full-History	9.0	53.5	61.5	4.5
	BM25	13.5	49.5	49.5	1.0
	Dense RAG	22.5	6.0	9.0	0.5
	Nano-Mem	23.0	12.5	18.0	0.0
	A-MEM	15.5	26.0	30.5	0.5
	Mem0	22.0	7.5	9.5	0.0
	LD-Agent	25.6	23.5	19.0	0.5
	Pact-Mem	**48.0**	**54.0**	**54.5**	**26.5**
	Ref-Traj	33.5	88.0	90.0	38.0

Table 16: Capability pass rates (%) by formation mechanism under Ref-Traj with DeepSeek V4 Flash.

	User-compression setting				Agent-compression setting			
Mechanism	Acpt@S_3	Acpt@S_4	Acpt@S_5	Prod@S_6	Cmpr@S_3	Acpt@S_4	Acpt@S_5	Prod@S_6
Inversion	86.0	86.0	74.0	22.0	22.0	94.0	98.0	38.0
Anchoring	74.0	88.0	80.0	8.0	18.0	80.0	82.0	50.0
Adoption	78.0	80.0	72.0	30.0	34.0	84.0	84.0	26.0
Erosion	72.0	60.0	72.0	24.0	60.0	94.0	96.0	38.0

E.3 Agent-compression retrieval diagnostics

Table 17 applies the provenance-based retrieval diagnostics from Table 4 to agent compression. Repeated form use raises intermediate recall, but joint M-and-F hits at S_6 remain 2.5%–11.0% across memory systems.

Table 17: Retrieval diagnostics (%) in the agent-compression setting. Recall measures formation-session coverage in the top-five retrieval. Hit@S_6 reports retrieval of an M-establishing session, an F-using session, or both. Column maxima are bold.

(a) Recall by memory system

Method	S_3	S_4	S_5	S_6
BM25	**44.5**	**47.2**	**52.4**	12.4
Dense RAG	20.5	10.0	21.6	5.2
Nano-Mem	21.2	10.7	22.8	5.2
A-MEM	34.8	15.7	26.0	10.9
Mem0	16.2	11.8	18.0	10.0
LD-Agent	42.2	37.2	35.1	**18.2**

(b) Hit@S_6

Method	M	F	M & F
BM25	11.0	46.0	7.0
Dense RAG	12.5	10.0	2.5
Nano-Mem	12.5	10.0	2.5
A-MEM	10.0	34.0	3.5
Mem0	11.5	32.0	4.5
LD-Agent	**18.5**	**56.5**	**11.0**

E.4 Retrieval by formation mechanism

Table 18 separates BM25 recall by formation mechanism. All four mechanisms lose recall by S_6 in both compression settings despite their different form and meaning trajectories.

Table 18: BM25 formation-session recall (%) by formation mechanism and setting.

	User compression				Agent compression			
Mechanism	S_3	S_4	S_5	S_6	S_3	S_4	S_5	S_6
Adoption	**82.0**	40.0	**50.5**	**15.6**	49.0	38.7	56.5	10.0
Anchoring	40.0	**52.0**	28.0	6.8	18.0	41.3	43.5	12.4
Erosion	75.0	45.3	43.0	10.4	46.0	40.0	46.0	10.8
Inversion	70.0	50.7	42.5	12.4	**65.0**	**68.7**	**63.5**	**16.4**

E.5 Detailed content-level diagnostics

Table 19 expands the joint-evidence results in Table 7 by reporting separate form and private-meaning retrieval. Store-level availability is reported only for user compression because the saved agent-compression outputs expose retrieved context but not the complete memory store.

Table 19: Detailed content diagnostics at S_6 with DeepSeek V4 Flash (%). F is the exact convention form and M is at least one private-meaning anchor. The final column reports production among cases retrieving both, with counts in parentheses.

(a) User-compression setting

Method	Store F & M	Ret. F	Ret. M	Ret. F & M	Prod | F & M
BM25	**99.0**	44.5	67.5	38.5	1.3 (1/77)
Dense RAG	**99.0**	38.0	57.5	29.0	0.0 (0/58)
Nano-Mem	**99.0**	6.0	19.0	5.5	0.0 (0/11)
A-MEM	**99.0**	35.0	57.0	26.5	0.0 (0/53)
Mem0	64.5	5.5	11.0	3.0	16.7 (1/6)
LD-Agent	60.0	19.0	34.0	9.0	11.1 (2/18)
Pact-Mem	83.5	**76.0**	**80.5**	**75.5**	**17.2** (26/151)

(b) Agent-compression setting

Method	Ret. F	Ret. M	Ret. F & M	Prod | F & M
BM25	47.0	63.0	31.5	1.6 (1/63)
Dense RAG	5.5	14.0	5.5	9.1 (1/11)
Nano-Mem	6.0	13.5	4.5	0.0 (0/9)
A-MEM	39.0	60.5	31.0	1.6 (1/62)
Mem0	7.0	13.5	5.0	0.0 (0/10)
LD-Agent	32.5	34.5	15.5	3.2 (1/31)
Pact-Mem	**70.5**	**76.0**	**66.0**	**40.2** (53/132)

F PACT-MEM IMPLEMENTATION

F.1 IMPLEMENTATION DETAILS

Joint form–meaning storage. Each memory entry stores the exact expression together with its private meaning, usage conditions and grounded dialogue evidence. The extractor reads the latest 12 closed sessions and never receives the registered form, meaning or reference reply.

Cross-session evidence consolidation. Repeated uses update one convention-level entry rather than producing independent session summaries. A recurrence-guided recovery pass revisits phrases that appeared across closed sessions when ordinary extraction misses the binding. The memory model must infer the meaning from the observed dialogue; recurrence alone cannot create an entry.

Form-free situational retrieval. Each entry stores views of the situations in which the convention was used, with words from the form removed. Dense retrieval matches the current opening against these views and returns at most 15 complete form–meaning entries with their evidence. Retrieval views are never shown to the backbone LLM, and the returned memory does not instruct it to use the form.

All three components operate only on closed history. Pact-Mem adds no backbone-LLM instruction beyond the shared prompt above. Benchmark annotations support evaluation and provenance logging but do not provide the registered convention or reference reply to the extractor, retriever or backbone LLM.

G BEHAVIORAL CONTROL FOR PACT-MEM

Section 5.3 shows that exact formation evidence does not ensure production. We therefore compare Pact-Mem with **Pact-Mem + Convention-Use Guidance**, which instructs the backbone LLM to use an established dyadic convention when appropriate while leaving storage, consolidation and retrieval unchanged. This intervention separates memory access from the backbone LLM's tendency to realize a retrieved convention; Table 20 reports the result.

Table 20: DeepSeek V4 Flash $\mathrm{Prod@}S_6$ pass rates (%) with and without convention-use guidance. Each cell contains 200 examples. Ref-Traj supplies the exact preceding reference formation sessions. Bold marks the better Pact-Mem condition.

Context condition	User compression	Agent compression
Pact-Mem	13.0	26.5
Pact-Mem + Convention-Use Guidance	**14.5**	**28.5**
Ref-Traj	21.0	38.0

Because the first two rows use the same memory pipeline, their difference isolates the instruction. Convention-use guidance raises production by 1.5% under user compression and 2.0% under agent compression. Both scores remain below Ref-Traj, which reaches 21.0% and 38.0%. A direct instruction changes realization only slightly and does not close the gap to exact formation evidence. This result supports treating response behavior as a separate limitation rather than an omitted prompt clause.

H Related work

This appendix expands Section 2 around the three distinctions that define PactBench: stated versus induced memory, public versus partner-specific grounding, and knowing versus using a convention.

H.1 Long-term conversational memory

Long-term conversation benchmarks query information stated in multi-session history (Maharana et al., 2024; Wu et al., 2025; Tan et al., 2025a; Jiang et al., 2025; Wu et al., 2026a). Memory systems compress that history into summaries, facts, linked notes, segments, or graphs before retrieval (Zhong et al., 2023; Chhikara et al., 2025; Xu et al., 2025; Pan et al., 2025; Tan et al., 2025b; Rasmussen et al., 2025; Yue et al., 2026; Li et al., 2025b; Wang & Chen, 2025). LoCoMo-Plus and T-Mem test low-overlap retrieval, but their target remains a constraint expressed by an earlier cue (Li et al., 2026; Guo et al., 2026b).

PactBench instead forbids any turn from stating the form–meaning mapping. Success therefore requires induction across sessions rather than retrieval of a sentence. Ref-Traj isolates this memory requirement from response generation. MAGNIFICo also tests the application of novel interpretations supplied in context, but not their collaborative formation or retention across sessions (Patel et al., 2023).

Role-play, companionship, and personalized-agent benchmarks evaluate persona consistency, social support, or recall of user-provided information (Wang et al., 2024; Liu et al., 2026; Li et al., 2025a; Wu et al., 2026d;a). Their target material is authored before interaction or supplied by the user, whereas a dyadic convention is jointly formed and partner-specific.

H.2 Convention formation and partner-specific grounding

Conceptual pacts are temporary, partner-specific agreements established through use (Brennan & Clark, 1996). Repeated interaction shortens referring expressions, and the result depends on the pair that formed it (Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996). Current models often understand a partner's shortening without producing comparable contraction or partner-specific adaptation (Hua & Artzi, 2024; Wang et al., 2026; Jones et al., 2026; Poelitz et al., 2026). Post-training or cost-sensitive self-play can strengthen this behavior (Hua et al., 2026; Vaduguru et al., 2025).

Related benchmarks either provide a novel word's definition or let symmetric agents converge on a public lexicon (Eisenschlos et al., 2023; Patel et al., 2023; Vera, 2026). These settings are usually single-session, expose the meaning directly or publicly, and evaluate repeated reference rather than distinct formation stages. PactBench instead distributes a private mapping across sessions and evaluates compression, acceptance, and production independently.

H.3 Pragmatics and knowing versus acting

Know-versus-act, deferred-intention, and preference-following benchmarks distinguish stored information from behavior (Feng et al., 2026; Liu & Gabriel, 2026; Fu et al., 2026; Wu et al., 2026b; Cai et al., 2026; Xiao et al., 2026; Guo et al., 2026a; Wu et al., 2026d). Their target is an explicitly stated fact, preference, or intention. PactBench instead requires an induced form–meaning binding and separately measures recovery, consistent response, and production.

Pragmatic benchmarks test irony, metaphor, indirect requests, and implicatures whose non-literal meanings are publicly inferable from local context (Hu et al., 2023; Yerukola et al., 2024; Yu et al., 2026; Arai & Ren, 2026). PactBench requires a partner-specific mapping learned from prior interaction. Unlike conflicts between external context and parametric facts (Xu et al., 2024), neither reading is factually privileged; the dyad determines the convention.

The evaluation protocol is itself a documented source of distortion. Disclosing the task type in the prompt shifts scores unevenly across task categories, and string-overlap metrics track generation length rather than semantic correctness (Li et al., 2026). Our probes carry no task label and judge an admissible response set rather than overlap with one reference string. Appendix C.2 audits DeepSeek V4 Pro against a human reference annotator across both compression settings.