SCOPE: MEMORY CONFLICT AS A QUESTION-TIME RELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conversational memory accumulates statements that revise, qualify, or contradict earlier ones. Which statement matters depends on a question that may not yet have been asked. Resolving conflicts when memory is written can discard a distinction a later question needs. We ask what memory should decide before that question arrives. Keeping the dialogue preserves these distinctions, but a bounded reader still needs to find and select the evidence relevant to each question. We introduce SCOPE (Slot-indexed Conversational Organization for Provenance-linked Evidence), which groups extracted statements by subject and attribute while retaining links to their original dialogue. At query time, these groups guide access to source turns for an answering model to interpret. On MemConflict, a benchmark of questions over conflicting conversations, SCOPE answers more accurately overall and on questions about change over time than lexical retrieval with the same evidence selection and answering steps. It also exceeds the evaluated memory systems in overall accuracy; both results hold under a second LLM judge. These findings support separating memory organization from question-dependent interpretation of competing evidence.

1 INTRODUCTION

A conversational agent may learn that its user works at North, later joins South, and would consider West if an offer arrives. All three statements concern the same attribute, yet questions about the current employer, an employment change, and a conditional plan require different evidence. The earlier statement remains useful after the move to South, while West remains a conditional possibility. A memory that stores only the current value answers the first question and cannot answer the other two. Memory conflict therefore raises a design question: *what should memory decide before the question arrives?*

Memory systems can fail a later question in two ways. Agent memories organize experience through reflection, tiered storage, extraction and consolidation, or linked notes (Park et al., 2023; Packer et al., 2023; Chhikara et al., 2025; Xu et al., 2025; Li et al., 2025); conflict-aware designs go further and judge conflict or validity over stored claims, revoking, invalidating, or resolving them (Zhou et al., 2026; Rasmussen et al., 2025; Volkov et al., 2026). Consolidating into one accessible value can remove a distinction a future question needs, while retrieving an isolated mention can omit the exchange that qualifies it. Recent benchmarks make these failures measurable through conflicting statements, temporal updates, implicit invalidation, and nuanced relations among memories (Tao et al., 2026; Wu et al., 2025; Hu et al., 2026; Chao et al., 2026; Wang et al., 2026), and through answering with hallucinated, insufficient, or conflicting evidence (Chen et al., 2025; Yang et al., 2026b;a).

The first challenge can be formalized through an answer-preservation condition: histories mapped to the same memory state must admit a common correct answer to every future question (Equation 4). Retaining the full dialogue avoids irreversible storage loss but moves the difficulty to retrieval. A fixed context window cannot cover an indefinitely growing history, so a question must still reach the turns it needs, and the reader sees only a bounded selection of them; even within the window, models use long inputs unevenly (Liu et al., 2024).

We separate *membership*—what an assertion concerns—from *applicability*—whether and how it bears on a question. *Ingestion* processes dialogue into memory, checking source support and assign-

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.

ing membership without choosing a winning value. Assertions about the same subject and attribute share a *slot* without having to agree in value, so recording a move to South need not erase employment at North or turn the conditional plan for West into a current fact. SCOPE uses these source-linked slots as an index into the original dialogue. At query time, brief source excerpts (*witnesses*) help an LLM mapper identify relevant slots; links from their members recover the source sessions, from which SCOPE selects whole turns within the reader's budget. Extraction and slot assignment may omit or misgroup assertions, but they never delete the dialogue. *Assertions organize access; original dialogue supplies context; the question determines applicability.*

Deferring applicability creates two query-time tasks. *Access* identifies which sets bear on the question and recovers their linked source turns. *Allocation* selects from those turns under a bounded evidence budget. Keeping the dialogue preserves possible answers; access and allocation determine which of them the reader can use.

On MemConflict, a controlled intervention tests whether this organization changes answers: replacing slot access with lexical retrieval while keeping selection and the reader fixed lowers answer accuracy (AA) by 5.91 points over all questions and by 8.03 on dynamic questions. SCOPE also exceeds the evaluated memory systems under two LLM judges. Alongside the dynamic access gain, further interventions find static gains from retaining index occurrences and conditional gains from witnessed mapping. Evidence-budget controls distinguish retaining alternatives from selecting them for the reader.

Our contributions are:

- **A set-based formulation of memory conflict.** We separate subject–attribute membership from question-dependent applicability, representing competing assertions without requiring a single winning value. An answer-preservation criterion specifies which distinctions must remain accessible to future questions (Section 2).
- **Witnessed access from assertion sets to dialogue.** SCOPE uses source excerpts to identify relevant sets, expands their members back to original dialogue, and selects evidence for a bounded reader. Extracted assertions organize access; original turns supply the evidence (Section 3).
- **Evidence for separating retention from presentation.** A matched intervention shows overall and dynamic gains from slot access. Separate controls find static gains from index retention and replicated conditional gains from witnessed mapping. Under tighter excerpt budgets, keeping all occurrences eligible for selection can instead hurt dynamic accuracy—distinguishing retention from evidence selection (Sections 5.2 and 5.3).

2 A SET-BASED FORMULATION OF MEMORY CONFLICT

Memory conflict creates a timing problem: memory is built before a question arrives, but whether an assertion applies must be judged in light of that question. We distinguish the observed dialogue $H_{\leq t}$, an index I_t built before q that groups retained assertion occurrences $C_{\leq t}$ into slots, and question-specific evidence $E_q \subseteq H_{\leq t}$ supplied to a fixed reader. Source identifiers tie the index and evidence to the same turns and preserve their order.

2.1 OBSERVED HISTORY AND THE ANSWERING TASK

A source turn $u_i = (\iota_i, s_i, r_i, d_i, x_i)$ records its ordered identifier, session, speaker, supplied date, and text. The observable prefix $H_{\leq t}$ comprises the turns of sessions $s \leq t$, with $H_s = \{u_i : s_i = s\}$. The session cutoff t limits observation; it is distinct from the event time requested by a question. An evidence-access rule π uses memory $\mathcal{M}_t$ constructed from this prefix and supplies original turns to a fixed reader R:

$$E_q = \pi(q, \mathcal{M}_t), \qquad \hat{y}_q = R(q, E_q), \qquad E_q \subseteq H_{\leq t}, \quad \mathrm{bytes}(E_q) \leq B_E. \tag{1}$$

The objective is to answer accurately within the evidence budget B_E; $\mathrm{bytes}(E)$ measures the serialized UTF-8 JSON representation of the turn records, ordered by source identifier. Reference answers are used to evaluate $\hat{y}_q$; they are not inputs to π or R.

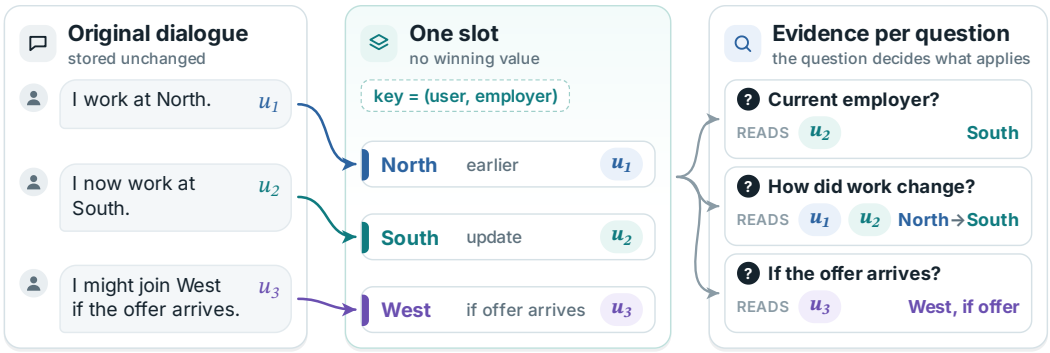


Figure 1: **Shared membership, question-dependent evidence.** Assertions extracted from three dialogue turns share an employer slot while retaining their source links and qualifications. Different questions require different source turns, and joining West remains a conditional possibility. Answers are illustrative.

2.2 Assertion occurrences and slots

An assertion occurrence has the representation

$$c = (\eta, e, a, v, \kappa, \tau, s, \Lambda), \tag{2}$$

where η identifies the occurrence, e its subject, a its proposed attribute name and definition, v its value, κ its condition, τ its temporal role, s its source session, and Λ its quoted source links and support relations. Multiple links may connect an assistant's proposal to a user's confirmation; the primary, value-bearing quoted link is the occurrence's witness. Occurrence identity keeps repeated values distinct, preserving the source of each assertion.

Let $C_{\leq t}$ contain retained occurrences from sessions $s \leq t$. Each occurrence has an assigned canonical or provisional attribute key $b(c)$, distinct from its proposed description $a(c)$; the key carries its stored definition. Membership is equality of assigned subject–attribute keys:

$$\begin{aligned} k(c) &= (e(c), b(c)), & c_i \sim c_j &\iff k(c_i) = k(c_j), \\ G_k(t) &= \{c \in C_{\leq t} : k(c) = k\}, & I_t &= \{(k, G_k(t)) : G_k(t) \neq \varnothing\}. \end{aligned} \tag{3}$$

Equality of assigned keys makes $\sim$ an equivalence relation whose classes partition $C_{\leq t}$. We call $G_k(t)$ an assertion set and $(k, G_k(t))$ a slot; $I_t^{\text{keys}} = \{k : G_k(t) \neq \varnothing\}$ denotes the index keys. Shared membership implies neither value agreement, $c_i \sim c_j \not\Rightarrow v(c_i) = v(c_j)$, nor applicability to the same question. Values, conditions, and temporal roles remain properties of individual occurrences; grouping does not collapse them into a representative value.

2.3 Competition and question-dependent applicability

Membership records *what an assertion concerns*; applicability asks *whether and how it contributes* to an answer. Assertions compete when distinct values bear on the same part of a question and the surrounding dialogue determines how they should be interpreted. We call this question-dependent case a *memory conflict*. Competition need not mean contradiction: earlier and current employers can both be valid and both be needed to describe a change. Conditional assertions must be interpreted under their stated conditions, including when a question asks about a hypothetical case.

An answer-preservation criterion. Consider histories North→South and East→South. Storing only the current employer maps both to South, yet "Which employer did the user leave?" requires different answers. Let $\mathcal{H}$ be a finite set of histories, $\mathcal{Q}$ a question set, $\mathcal{M}(H)$ all history-dependent information accessible to an answering rule, and $\mathcal{Y}^*(q, H)$ the nonempty set of acceptable answers. A rule using only $(q, \mathcal{M}(H))$ exists that is correct for every $H \in \mathcal{H}$ and $q \in \mathcal{Q}$ exactly when, for every reachable state m and question $q \in \mathcal{Q}$,

$$\bigcap_{H \in \mathcal{H}:\, \mathcal{M}(H) = m} \mathcal{Y}^*(q, H) \neq \varnothing. \tag{4}$$

The condition says that every history collapsed into the same memory state must admit one common answer to q (Appendix A). It constrains what a representation may erase; it does not certify the reader. A current-value view can meet it when an archive remains accessible. In SCOPE, the original dialogue remains stored alongside the index, while applicability is decided once a question arrives.

2.4 Query access and source expansion

Our memory representation is $\mathcal{M}_t = (H_{\le t}, C_{\le t}, I_t)$, with the slots I_t derived from the retained occurrences and their stored key assignments. Query access selects keys and takes their union:

$$K_q \subseteq I_t^{\text{keys}}, \qquad A_q(t) = \bigcup_{k \in K_q} G_k(t). \tag{5}$$

We write A_q for $A_q(t)$; it holds candidate occurrences, not accepted answers. Its source expansion is

$$\Sigma_t(A) = \bigcup_{c \in A} H_{s(c)} \quad (A \subseteq C_{\le t}), \qquad E_q \subseteq P_q \subseteq \Sigma_t(A_q) \subseteq H_{\le t}, \tag{6}$$

where P_q is the candidate pool drawn from the expansion under the byte budget B_P. Different groups may link to the same session; the union includes each turn only once.

Where the difficulty remains. Storing the whole history preserves source distinctions, but the reader receives only E_q, bounded by B_E. The same criterion applies to histories that the evidence rule maps to identical E_q: source retention cannot repair an uninformative reader input. Selection has the question in hand but relies on an index built without it. This raises two empirical questions: *access*, whether the question reaches retained dialogue through the index; and *allocation*, how the effect of presenting alternatives changes under an evidence budget. Sections 5.2 and 5.3 measure both.

3 SCOPE: from assertion sets to source evidence

SCOPE implements the formulation in three stages: ingestion constructs assertion sets, access maps a question to those sets, and allocation selects source turns for the reader. The sets organize *where to look*; original dialogue supplies *what to interpret*. LLMs handle paraphrase, attribution, conditions, and question-to-attribute matching without choosing a winning value or deleting dialogue. Figure 2 summarizes the procedure; Algorithm 1 and the decision contracts (Appendix B) give the details.

3.1 Ingestion: fallible, non-destructive organization

We construct *assertion occurrences* first and assign membership second, separating source support from the decision about which statements concern the same subject and attribute.

Retain the assertion and its support. A value may be proposed in one speaker's turn and confirmed or qualified in another, so extraction retains these links. For session H_s, extraction proposes assertion occurrences and Validate_s checks fields, quoted spans, and linked support:

$$Q_s = \text{Validate}_s(\text{Extract}(H_s; n)), \qquad |Q_s| \le n. \tag{7}$$

Each $c \in Q_s$ retains the fields of Equation 2 and its source links $\Lambda(c)$. When extraction yields more than $n = 32$ candidates, an LLM selects n of their IDs before source and support validation; the others never enter the index, although their turns remain in the dialogue. Assistant-originated values require a specific user confirmation or adoption span and an additional LLM support review; adopting a plan does not establish that the event occurred. These checks concern support and attribution, not which value is true.

Assign membership; retain distinct members. Membership depends on subject and attribute rather than value, so assertions with different values can share a group. Given the registry Γ of attributes accepted in earlier sessions, an LLM binds each proposal to an existing attribute from a lexical shortlist or proposes a new attribute. Let F_s contain occurrences with accepted bindings and

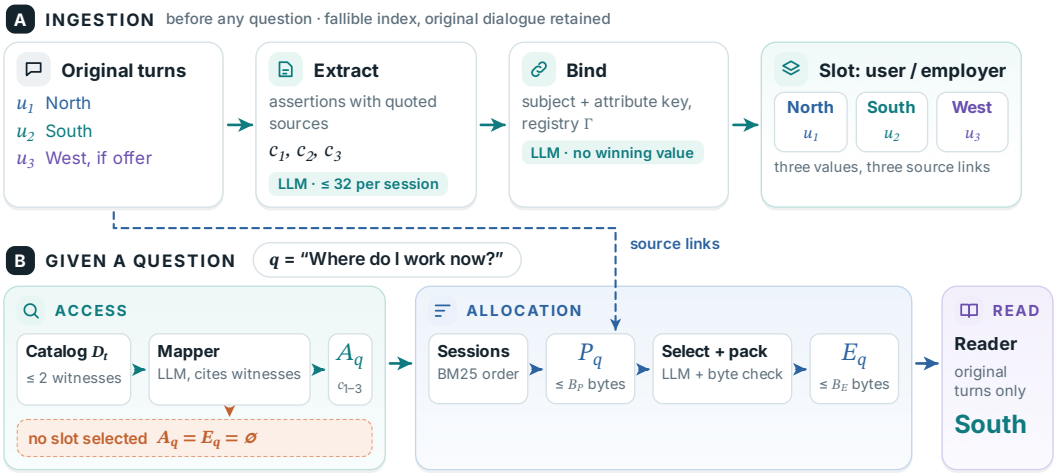


Figure 2: **From source-linked sets to bounded dialogue evidence.** Ingestion keeps original turns while indexing assertion occurrences. Given a question, witnessed mapping produces A_q (access); linked-session assembly and selection produce budgeted P_q and E_q (allocation). The dashed orange branch shows the possible empty-mapping path. Source links return to retained dialogue, so the reader receives original turns rather than extracted fields. The North/South/West answer is illustrative.

U_s validated occurrences whose binding remains unresolved; each member of U_s keeps a stable provisional key derived from its own proposed attribute. Insertion is a union update, $G_k \leftarrow G_k \cup \{c \in F_s \cup U_s : k(c) = k\}$, so earlier employment at North, a move to South, and a conditional plan for West can share the employer key while retaining their qualifications (Figure 1). Groups grow as sessions arrive, and earlier bindings are never revisited. Their composition therefore depends on ingestion order. Construction uses no evaluation questions or reference answers.

3.2 Access: From the Question to Groups

Witnesses identify a group; membership determines what its expansion reaches. For each cutoff-eligible group, the catalog D_t contains its key, definition, and at most two source witnesses: the first and last in saved occurrence order. Intermediate members remain stored but are not displayed. A question may describe an attribute differently from its stored name, so the mapper is instructed to judge relevance from the definition and witness content. Let W_k be the identifiers of the offered witnesses. With mapping limit m, the LLM mapper proposes keys with supporting witnesses; validated pairs form a partial function M_q, and each selected key must cite a nonempty subset of its offered witnesses:

$$(k, W_{q,k}) \in M_q \;\Rightarrow\; k \in I_t^{\text{keys}},\;\; \varnothing \neq W_{q,k} \subseteq W_k; \qquad K_q = \{k : (k, W_{q,k}) \in M_q\},\;\; |K_q| \leq m. \tag{8}$$

The selected keys define A_q through Equation 5, and all eligible members enter A_q, including those absent from the catalog. The mapper is instructed to abstain when no entry bears on the question; $K_q = \varnothing$ then gives $A_q = \varnothing$, and the primary configuration has no fallback. Mapping selects where to look; it does not decide which value answers q.

3.3 Allocation: From Groups to Bounded Evidence

Because a qualification may appear anywhere in the session where a value was stated, SCOPE expands every occurrence in A_q to its whole source session, $\Sigma_t(A_q)$ (Equation 6), rather than returning extracted values. Two consequences follow: a statement missing from the index can still reach the reader when another member links its session, and a confirmation in an unlinked session cannot. Let $S_q = \{s(c) : c \in A_q\}$ be the linked sessions and L_t the BM25 ranking of all eligible sessions for q (Robertson & Zaragoza, 2009). The candidate pool is

$$L_q = [\, s \in L_t : s \in S_q \,], \qquad P_q = \text{Assemble}_{B_P}(A_q, L_q; H_{\leq t}), \qquad \text{bytes}(P_q) \leq B_P. \tag{9}$$

BM25 orders the linked sessions for assembly without adding new ones. This ordering determines which turns are considered first under B_P. $\mathrm{Assemble}_{B_P}$ prioritizes primary source turns of all members of A_q, then adjacent and remaining turns, keeping whole turns (Appendix B.1).

Next, an LLM selector returns a validated priority list J_q of unique turn identifiers from P_q. The deterministic Pack_{B_E} packs whole turns in that order, skips those that do not fit, and presents the retained turns chronologically:

$$\begin{aligned} &E_q = \mathrm{Pack}_{B_E}(P_q, J_q), && \hat{y}_q = R(q, E_q), \\ &E_q \subseteq P_q \subseteq \Sigma_t(A_q) \subseteq H_{\le t}, && \mathrm{bytes}(E_q) \le B_E. \end{aligned} \tag{10}$$

On the access sample (Section 4), selection rather than the byte cap sets how much evidence the reader sees. The selector is instructed to retain relevant alternatives, revisions, and qualifications; delivered evidence averages 2,068 bytes, and the 24,000-byte cap removes no selected turns (Appendix F.3). We study tighter caps in the retention controls (Section 5.3). The reader receives original turns rather than extracted values, conditions, or temporal roles.

4 Experimental design

We evaluate all 3,750 MemConflict questions across 30 personas: 2,946 dynamic, 360 static, and 444 conditional (Tao et al., 2026). Dynamic questions concern changes over time; static questions ask which report of a stable fact is correct; conditional questions ask whether an assertion applies under specified conditions.

We compare SCOPE with BM25 (Robertson & Zaragoza, 2009), Mem0 (Chhikara et al., 2025), MemOS (Li et al., 2025), LangMem (LangChain, 2025), and Letta's released archival memory adapter (Packer et al., 2023). All methods receive the same dialogue prefixes with speaker roles, dates, and order; none receives structured persona profiles, and reference answers are used only for evaluation. Every method uses DeepSeek-V4.1-Flash at temperature zero with reasoning disabled, the released reader prompt, and the released judge. Memory construction and formatting remain method-specific. SCOPE maps at most $m = 3$ groups into a $B_P = 48{,}000$-byte pool. SCOPE and BM25 share a $B_E = 24{,}000$-byte evidence budget; BM25 passes its top-ranked chunks directly to the reader (*direct BM25*).

Mechanism analyses use a fixed 360-question *access sample*: eight dynamic, two static, and two conditional questions per persona. A full-prefix reader is evaluated on this sample, with median prompts of 206,285 proxy tokens (Appendix D.3). Answer accuracy (AA) is the mean of the released judge's per-question scores, which include partial credit on dynamic and static questions. Dynamic AA is the primary endpoint. Secondary metrics are pooled AA over all questions and equal-weight type AA, which averages dynamic, static, and conditional AA with one-third weight each, regardless of their question counts. Pipeline and evaluator failures count as zero. Paired intervals bootstrap the represented personas (30 in full-cohort comparisons), with 20,000 resamples and seed zero (Efron & Tibshirani, 1993). Appendix D details settings, uncertainty, and recorded model provenance.

5 Results on memory conflict

5.1 Full-benchmark comparison

SCOPE leads the six methods in pooled AA (52.07% over 3,750 questions) and dynamic AA (50.32% over 2,946; Table 1). The gains over BM25 are 13.04 points pooled and 16.45 dynamic (Appendix D.1). SCOPE trails on conditional questions (77.03% versus 83.11%); Section 5.2 locates this deficit in empty access. Re-judging with a second model family (Qwen3.8-Flash) preserves the pooled ranking and paired gains (Appendix D.4); Appendix D.2 tests LangMem's retrieval width. On the access sample, full-prefix reading scores 53.61% versus SCOPE's 52.50% (paired $+1.11$ $[-3.06, +5.42]$ points), establishing neither advantage nor equivalence (Appendix D.3).

Transfer to MemoryAgentBench. With extraction adapted to numbered facts (Hu et al., 2026), SCOPE leads the evaluated methods in single-hop conflict resolution at four context sizes (6K–262K; Table 10; Appendix E). On the same 6K and 64K histories, multi-hop accuracy falls from 100% and 95% to 25% and 21% for SCOPE, and from 99% and 91% to 10% and 5% for BM25.

Table 1: **Full MemConflict AA under dialogue-only input (%).** No structured persona profiles are supplied to any method. Common backbone, released reader prompt and AA judge; 3,750 questions and 30 personas. BM25 uses whole-turn chunks. LangMem includes five pipeline errors as zero. Bold marks the best score in each column.

Method	Pooled	Dynamic	Static	Conditional	Equal-weight type AA
SCOPE	**52.07**	**50.32**	**35.56**	77.03	**54.30**
BM25	39.03	33.88	26.81	**83.11**	47.93
Mem0	33.40	30.74	19.72	62.16	37.54
LangMem	37.17	37.54	30.83	39.86	36.08
MemOS	34.48	32.16	29.03	54.28	38.49
Letta	32.73	35.39	14.44	29.95	26.60

Table 2: **Full-cohort access ablation.** AA (%) on the same 3,750 questions and 30 personas; Δ is SCOPE minus no-slot access in percentage points, with paired 95% persona-bootstrap intervals. Both arms retain the same selector and reader.

Type	n	SCOPE	No-slot	Δ	95% CI
Pooled	3750	52.07	46.16	$+5.91$	$[+3.60, +8.03]$
Dynamic	2946	50.32	42.29	$+8.03$	$[+5.31, +10.61]$
Static	360	35.56	28.89	$+6.67$	$[+2.50, +10.83]$
Conditional	444	77.03	85.81	-8.78	$[-12.05, -5.46]$

5.2 Access: slot-access gains and empty mappings

We test slot access through a controlled intervention, then trace which sources reach the reader.

Matched access intervention. Slot access improves accuracy when every later stage is held fixed. The no-slot control replaces mapping and source expansion with whole-turn lexical retrieval on all 3,750 questions. It keeps the selector, released reader and judge, cutoffs, and byte budgets fixed. SCOPE improves pooled AA from 46.16% to 52.07%, a paired gain of 5.91 points [3.60, 8.03], and dynamic AA by 8.03 points [5.31, 10.61]. Static AA also improves; conditional AA falls by 8.78 points (Table 2). The Qwen judge also retains the pooled and dynamic gains (Appendix D.4). This tests the integrated access path, not its components separately. Appendices F.1 and F.4 detail response reuse, failures, and a realized-size control. Figure 3 shows persona-level gains.

On the 360-question access sample, including empty cases, mean reader evidence is 2,068 bytes for SCOPE versus 4,971 for no-slot access. SCOPE delivers turns absent from the lexical candidate pool on 252 questions (70%; Appendix F.3).

Although 88.42% of groups at each persona's final cutoff are singletons, 2,336 of 3,750 questions map to a multi-occurrence group. Gains appear in both strata; within dynamic questions, which dominate the multi-occurrence stratum, they are 16.18 points for multi-occurrence and 13.83 for singleton-only mappings (Appendix F.6). These comparisons show where access gains occur, but not how much grouping itself contributes.

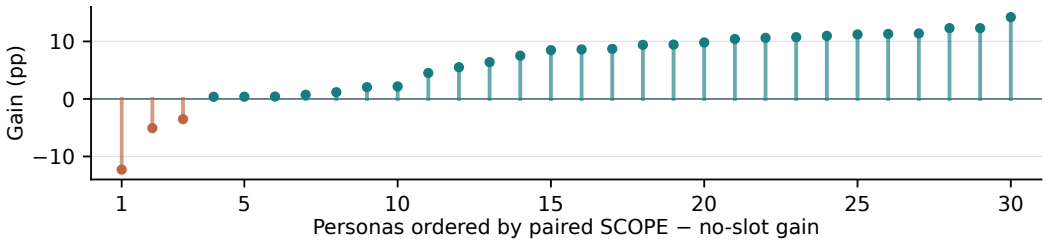


Figure 3: **Matched access gains by persona.** SCOPE minus no-slot AA, ordered by gain. Three of 30 are negative; the question-weighted pooled contrast is $+5.91$ points (Table 2).

Coverage. Empty mapping leaves SCOPE without evidence on 623 of 3,750 questions (16.61%), all scored zero; their catalogs are nonempty, but the mapper selects no slot. The accuracy loss is largest on conditional questions: direct BM25 correctly answers 67.24% of the 58 conditional empty-access questions, and these questions account arithmetically for SCOPE's aggregate conditional deficit against direct BM25 (Appendix F.5). In the full-prefix comparison, SCOPE leads on the 304 questions where it reaches evidence (62.17% versus 57.89%) but scores zero on the other 56, where full-prefix reading scores 30.36% (Appendix D.3). Figure 4 shows access outcomes.

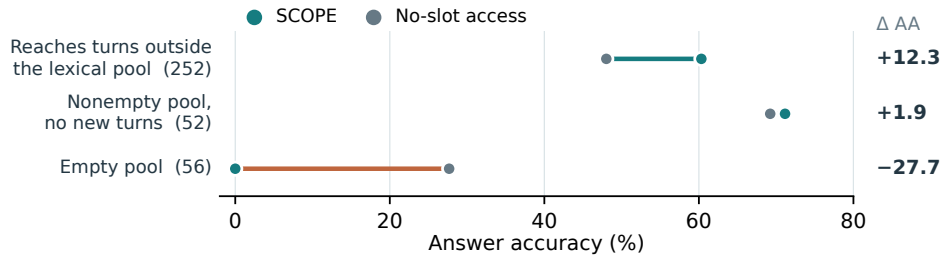


Figure 4: **Answer accuracy by SCOPE's source-access outcome.** Each row compares SCOPE and no-slot access on the same questions from the 360-question access sample (counts in parentheses). Δ is SCOPE minus no-slot AA in percentage points. Groups reflect SCOPE's retrieved evidence, not whether the evidence needed to answer was found.

Backup for empty access. Because original dialogue remains available, an empty mapping can trigger a second access route. When the candidate pool is empty—a condition observable before answering—we supply a lexical pool to the unchanged selector and reader. This post-hoc backup raises pooled AA from 52.07% to 57.23% and conditional AA from 77.03% to 87.39%, removing the conditional deficit against direct BM25. Other SCOPE answers remain unchanged; the primary results exclude the backup (Table 17; Appendix F.7).

Witnessed mapping. Source witnesses improve conditional-question accuracy across two disjoint question samples. The control maps questions using slot names and definitions, removing witness content and its citation requirement while leaving the index and downstream stages unchanged. Conditional AA improves by 26.67 and 30.00 points in the initial and confirmation samples, respectively, each containing 60 conditional questions. Across the two 360-question samples, conditional questions account for the 4.72-point pooled gain. The dynamic difference is $+0.31$ points on 480 questions, with an interval spanning zero (Appendix F.8).

5.3 Retention: index access and evidence allocation

What remains in the index. Retaining occurrences in the index matters even when the dialogue remains stored. On the 360-question access sample, we rerun the same mapper and downstream stages with all-occurrence or latest-only indexes. All-occurrence raises static AA from 20.00% to 39.17% but has a lower dynamic estimate than latest-only (48.96% versus 52.71%); both dynamic and pooled intervals include zero (Table 3). Holding the archive fixed separates what remains stored from how the pipeline reaches it. Session expansion can still recover statements whose index occurrences were removed (Appendix G).

Table 3: **All-occurrence versus latest-only index AA (%).** Same 360 questions and retained dialogue. Differences are SCOPE minus the latest-only index, with paired 95% persona-bootstrap intervals. Bold marks higher observed AA.

Type	n	SCOPE	Latest-only index	Δ [95% CI]
Sample pooled	360	**52.50**	50.69	$+1.81$ $[-2.22, +6.11]$
Dynamic	240	48.96	**52.71**	-3.75 $[-9.17, +1.87]$
Static	60	**39.17**	20.00	$+19.17$ $[+12.50, +25.83]$
Conditional	60	**80.00**	73.33	$+6.67$ $[+0.00, +15.00]$

What reaches the reader. With SCOPE's mapped slots fixed, we compare candidate excerpts from all occurrences with those from only the latest per slot, using quoted source spans rather than whole turns. Both pass through the same selector and reader. Under the standard cap, all-occurrence improves pooled AA by 6.81 points but delivers nearly twice as many bytes (729 versus 384). We then cap each question at its latest-only evidence size, either repacking selected excerpts or rerunning selection. Pooled gains shrink to 0.69 and 2.08 points, with both intervals spanning zero. The dynamic gain of 4.17 points reverses to losses of 7.08 and 4.37 points, while static and conditional gains remain. Equal caps still allow different delivered sizes (Table 19; Appendix G.4). These results distinguish preserving alternatives in memory from allocating evidence for an answer.

6 Related work

Conflict-aware memories differ in what they decide about stored claims and what they expose to the answering model. TEPA revokes keyed precedents (Zhou et al., 2026); Zep invalidates temporal graph edges while retaining source episodes (Rasmussen et al., 2025); and StateFuse explicitly represents conflicting claims for task-time resolution (Volkov et al., 2026). These systems retain history or provenance while representing claim validity or conflict. SCOPE gives extracted assertions a different role: their subject–attribute membership organizes access to original dialogue. StateFuse's (namespace, subject, predicate) key resembles this grouping, but its task-specific view returns selected claims or unresolved conflicts. In SCOPE, witnesses identify relevant slots, their members link back to source dialogue, and the reader interprets original turns; the index records membership without a conflict label or winning value. This separates organizing evidence from adjudicating its applicability, rather than moving the same resolution step to a later stage. Table 23 compares these interfaces; Appendix I covers further related work.

7 Discussion and limitations

The design tension is between future reuse and present relevance. An occurrence can remain an entry point to its source session even when its value is unnecessary for the current answer. The retention results support separating these decisions: memory preserves access to alternatives, while evidence selection determines what the reader sees for a particular question. Omitting a turn from the reader's context need not remove it from memory. Empty access also provides an observable signal for routing to lexical retrieval, full-prefix reading, or abstention (Yang et al., 2026a); only the lexical backup was tested. The evaluation uses one method backbone and two model-based judges. SCOPE requires more input tokens per question than BM25 (61,014 versus 5,898, excluding judging and including amortized ingestion). Appendices F.1 and G detail the measured access and retention limits.

8 Conclusion

SCOPE organizes assertions by what they concern without deciding which value should answer a future question. Source-linked assertion sets guide access, while original dialogue supplies evidence for question-dependent interpretation. On MemConflict, this access path improves overall and dynamic accuracy over lexical access with selection and reading fixed. Retaining index occurrences improves static accuracy, and witnessed mapping improves conditional accuracy across disjoint question samples. Under tighter excerpt budgets, however, keeping all occurrences eligible for selection can reduce dynamic accuracy. Together, these findings support separating what memory retains from what an answer requires: alternatives can remain available without entering every reader context. Memory conflict can thus be treated as a question-time relation among accessible sources, rather than a status that must be assigned to stored claims.

REPRODUCIBILITY STATEMENT

The supplementary material, also available through an anonymous repository at `https://anonymous.4open.science/r/scope-review-artifact-D762`, contains selected method source with documentation-only edits, evaluated prompt templates, question-level answers and scores for the six-method MemConflict comparison and full-cohort no-slot intervention, saved second-judge scores, and selected question-level measurements cited in the appendices. Edited files keep the evaluated executable syntax, and an offline check verifies integrity, accuracy, and paired persona intervals. Experiment orchestration, benchmark dialogue and references, native baseline packages, provider call records, and credentials are excluded; the supplement does not support a turnkey inference rerun. Runs use recorded model aliases, whose provider-side weights may change.

ETHICS STATEMENT

We evaluate released benchmark conversations and collect no new participant data. Source-linked memory can surface sensitive dialogue when used with real users; this study does not establish privacy safeguards or deployment reliability. Model judgments and unannotated index bindings should not be treated as verified facts about people.

AI USE STATEMENT

In this work, we used generative AI tools to help implement methods: AI coding assistants, following the authors' specifications, helped write code for SCOPE and for the scripts that generate tables and figures. We did not use generative AI tools for any other task requiring disclosure, including ideation, theory and proofs, experimental design, data generation or cleaning, and interpreting results. Additionally, we used generative AI tools to search for and summarize related work, format references, improve the readability of the writing, and create figures. We have reviewed all AI-assisted work: the authors reviewed all AI-assisted code and verified the generated tables and plots against experimental outputs, read every cited work and confirmed that each reference exists, and checked that edited text and figures accurately reflect our method and results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM Agents Know When Their Memories Are No Longer Valid? *arXiv preprint arXiv:2605.06527*, 2026.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. HaluMem: Evaluating Hallucinations in Memory Systems of Agents. *arXiv preprint arXiv:2511.03506*, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160.

Jon Doyle. A Truth Maintenance System. *Artificial Intelligence*, 12(3):231–272, 1979. doi: 10.1016/0004-3702(79)90008-0.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions. In *International Conference on Learning Representations (ICLR)*, 2026.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

LangChain. LangMem: SDK for Agent Long-Term Memory, 2025. URL https://langchain-ai.github.io/langmem/. Software documentation; accessed September 2026.

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. *arXiv preprint arXiv:2507.03724*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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.

George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. Blocking and Filtering Techniques for Entity Resolution: A Survey. *ACM Computing Surveys*, 53(2):31:1–31:42, 2020. doi: 10.1145/3377455.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22. ACM, 2023. doi: 10.1145/3586183.3606763.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. *arXiv preprint arXiv:2501.13956*, 2025.

Stephen 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.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. MemConflict: Evaluating Long-Term Memory Systems Under Memory Conflicts. *arXiv preprint arXiv:2605.20926*, 2026.

Sergey Volkov, Yang Li, and Ye Luo. StateFuse: Deterministic Conflict-Preserving Memory for Multi-Agent Systems. *arXiv preprint arXiv:2607.05844*, 2026.

Wenxuan Wang, Haoyu Sun, Fukuan Hou, Mingyang Song, Weinan Zhang, Yu Cheng, and Yang Yang. SubtleMemory: A Benchmark for Fine-Grained Relational Memory Discrimination in Long-Horizon AI Agents. *arXiv preprint arXiv:2606.05761*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *International Conference on Learning Representations (ICLR)*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Lu Yang, Shusheng Xu, Zhuoran Li, Tongkai Yang, and Longbo Huang. When Personal Memory Has No Single Answer: Evaluating LLM Agents under Irreducible Conflict. *arXiv preprint arXiv:2608.13921*, 2026a.

Tiancheng Yang, Matthias Schonlau, and Ilia Sucholutsky. Selective QA over Conflicting Multi-Source Personal Memory: A Diagnostic Testbed and Method Comparison. *arXiv preprint arXiv:2605.30087*, 2026b.

Yan Zhou, Yue Ouyang, Kaiyang Zheng, and Suncheng Xiang. TEPA: Revoking Stale Memories for Conflict-Robust Language Agents. *arXiv preprint arXiv:2608.07429*, 2026.

APPENDICES

The appendices explain the answer-preservation condition and implementation (Appendices A–C), evaluation protocol and comparison checks (Appendices D–E), and evidence for the access and retention analyses (Appendices F–G). Calls and tokens are in Appendix H; related work is in Appendix I.

A ANSWER-PRESERVATION CONDITION

For necessity, any answering rule $f(q, m)$ returns the same answer for every history with $\mathcal{M}(H) = m$. Correctness on all of them requires this answer to belong to the intersection in Equation 4. For sufficiency, choose one answer from each nonempty intersection and define $f(q, m)$ to return it; the rule is then correct for every history and question under consideration. This is an existence condition, without computational restrictions. The acceptable sets express semantic correctness, not the benchmark's partial-credit scores.

For a fixed question q, the same condition applies to question-specific evidence by replacing $\mathcal{M}(H)$ with $E_q(H) = \pi(q, \mathcal{M}(H))$. Histories that yield identical evidence give the reader the same input, so one answer must be acceptable for all of them. SCOPE does not guarantee this evidence-level condition: mapping and packing can omit necessary turns even though the full memory retains the original history.

B DETAILED ALGORITHM AND VALIDATION CONTRACTS

Algorithm 1 gives the complete procedure using the representation $\mathcal{M}_t = (H_{\leq t}, C_{\leq t}, I_t)$. Construction processes the observed prefix without the question. Accepted and provisionally keyed assertion occurrences are denoted by F_s and U_s throughout. Querying returns original source turns, not assertion summaries.

Operator contracts. EXTRACTLLM includes bounded schema repair and capacity selection before source and support checks. VALIDATEANDREVIEWSUPPORT checks schema fields, source ownership, exact quotations, and assistant-linked support. CATALOG uses saved occurrence order for first/last witnesses and orders entries by subject and assigned attribute; set notation suppresses this traversal metadata. VALIDATEMAP accepts at most m distinct offered catalog keys, each citing a nonempty subset of its offered witnesses (Equation 8). Only F_s updates Γ. Turn-ID validation rejects unknown or duplicate identifiers. The remaining operators follow Section 3; Table 4 shows where errors affect access or allocation.

Table 4: **Fallible decisions in the SCOPE pipeline.** Errors affect access or allocation, while the original dialogue remains stored.

Decision (made by)	Stage	Can be wrong about	Effect of an error
Extraction, capacity cap (LLM)	ingestion	which assertions exist	missing entry (access)
Support review (LLM)	ingestion	assistant attribution	missing entry (access)
Binding to Γ (LLM)	ingestion	an assertion's attribute	split/merged group (access)
Catalog witnesses (rule)	index	what the mapper sees	group missed (access)
Mapping (LLM)	query	which groups bear on q	empty or wrong A_q (access)
Candidate pool (BM25)	query	sessions that fit B_P	truncated pool (allocation)
Selection, packing (LLM)	query	turns shown to the reader	displaced evidence (allocation)

The algorithm shows the validated path. Empty catalogs skip mapping, and empty pools skip selection; valid empty evidence still reaches the reader. Malformed outputs are recorded separately from these valid empty cases. The following implementation details specify bounded repairs and resource limits.

Algorithm 1 SCOPE: memory construction and query-time evidence access

Require: Fixed reader R; proposal limit n; mapping limit m; byte budgets B_P, B_E

procedure ConstructMemory($H_{\leq t}$)
  $C \leftarrow \varnothing$; $\Gamma \leftarrow \varnothing$
  for each session H_s in chronological order, $s \leq t$ **do**
    $X_s \leftarrow \mathsf{ExtractLLM}(H_s, n)$
    $Q_s \leftarrow \mathsf{ValidateAndReviewSupport}(X_s, H_s, s)$
    $F_s \leftarrow \varnothing$; $U_s \leftarrow \varnothing$
    if $Q_s \neq \varnothing$ **then**
      $Z_s \leftarrow \mathsf{BindLLM}(Q_s, \Gamma)$
      $(F_s, U_s) \leftarrow \mathsf{AssignValidatedKeys}(Z_s, Q_s, \Gamma)$
    end if
    $C \leftarrow C \cup F_s \cup U_s$
    $\Gamma \leftarrow \mathsf{UpdateRegistry}(\Gamma, F_s)$
  end for
  $C_{\leq t} \leftarrow C$; $I_t \leftarrow \mathsf{GroupByKey}(C_{\leq t})$
  return $\mathcal{M}_t = (H_{\leq t}, C_{\leq t}, I_t)$
end procedure

procedure Answer($q, \mathcal{M}_t$)
  $D_t \leftarrow \mathsf{Catalog}(I_t)$ ▷ Key, definition, first/last witnesses
  $K_q \leftarrow \varnothing$; $M_q \leftarrow \varnothing$
  if $D_t \neq \varnothing$ **then**
    $Z_q \leftarrow \mathsf{MapSlotsLLM}(q, D_t, m)$
    $M_q \leftarrow \mathsf{ValidateMap}(Z_q, D_t, m)$
    $K_q \leftarrow \{k : (k, W_{q,k}) \in M_q\}$
  end if
  $A_q \leftarrow \bigcup_{k \in K_q} G_k(t)$; $S_q \leftarrow \{s(c) : c \in A_q\}$
  $L_t \leftarrow \mathsf{BM25Rank}(q, H_{\leq t})$
  $L_q \leftarrow [\, s \in L_t : s \in S_q \,]$ ▷ Preserve ranking; do not rescore
  $P_q \leftarrow \mathrm{Assemble}_{B_P}(A_q, L_q; H_{\leq t})$
  $J_q \leftarrow [\,]$
  if $P_q \neq \varnothing$ **then**
    $Z'_q \leftarrow \mathsf{SelectTurnsLLM}(q, P_q)$
    $J_q \leftarrow \mathsf{ValidateIDs}(Z'_q, P_q)$
  end if
  $E_q \leftarrow \mathrm{Pack}_{B_E}(P_q, J_q)$
  $\mathsf{CheckEvidence}(E_q, P_q, H_{\leq t}, B_E)$
  $\hat{y}_q \leftarrow R(q, E_q)$
  return $(\hat{y}_q, E_q)$
end procedure

B.1 IMPLEMENTATION SETTINGS AND PERSISTENCE

Ingestion failures. Missing extraction output can leave a session with no proposals, while absent or invalid binding decisions can leave anchored proposals unbound. These outcomes differ from successful extraction and are omitted from the algorithm's successful-path notation. Provisional keys are derived from the proposal's slot name and definition rather than from later sessions. The combined final evidence check in the algorithm summarizes validated identifier materialization plus the explicit cutoff and serialized-size checks.

Ranking and packing details. BM25 statistics are computed over all cutoff-eligible sessions before restricting the ranking to linked sessions; ties prefer the later session index. Assembly first interleaves primary source turns across linked sessions and slots, then considers neighboring and remaining turns. Each packing pass skips oversized units and continues, measuring serialized UTF-8 JSON including the empty array. Retained turns are presented in source order.

Frozen construction and access settings. Table 5 gives the output limits; Section 3 states the capacity cap $n = 32$ and first-and-last witness display, while Section 4 states the mapping cap $m = 3$. Capacity selection keeps exactly 32 original candidate IDs, preserving their text and links. The binding shortlist contains the six existing attributes with the highest lexical scores, computed from exact-name agreement, word-set overlap, and character-sequence similarity over names and definitions. Mapping, candidate-ID selection, and turn selection allow one validation-guided correction; these calls are retained and counted. Candidate and final evidence limits are $B_P = 48{,}000$ and $B_E = 24{,}000$ bytes. Appendix C gives the evaluated instructions and output contracts; Appendix E specifies the factual-memory transfer configuration.

Table 5: Main MemConflict configuration: maximum output tokens per LLM call. Capacity selection runs only for over-capacity extraction; native methods retain their own schemas.

Component	Output limit	Main implementation
Session extraction	12,000 / 24,000	Initial / schema repair
Capacity selection	600	Retain 32 original candidate IDs
Assistant support review	2,500	Validate linked assertion support
Registry binding	6,000	Existing equivalent attribute or new attribute
Slot mapping	1,000	At most three witnessed entries
Turn selection	2,500	Offered original IDs only
Reader	700	Released MemConflict prompt
Judge	1,000	Released AA prompt

Mechanical guarantees and their limits. Exact substring checks establish that a quoted span occurs in a supplied turn. They do not prove the span entails the proposed value, that a subject reference is correct, or that a tentative statement is an adopted plan. Proposals using only user spans bypass the assistant-support review; tentative user wording can therefore still be overinterpreted. Retaining a user confirmation link in the index also does not ensure that this link survives the final evidence budget. These are explicit failure modes, not hidden correctness assumptions.

C PROMPT TEMPLATES AND OUTPUT CONTRACTS

The evaluated instructions follow the three stages in Section 3: ingestion, access, and allocation. System messages are reproduced verbatim from completed runs. Each stage used one system instruction throughout; dialogue and other runtime inputs varied. We reproduce the witnessed-mapping instruction below. The ablation in Section 5.2 modifies its catalog fields and citation requirement. The exact messages for every stage, together with the repair templates, are included in the supplementary material.

Dynamic dialogue, registry candidates, questions, and catalogs are supplied as user-message data. Their layouts are described below; they are not additional natural-language instructions. Double-braced placeholders in the reader, judge, and repair templates mark runtime substitutions. No refer-

ence answers enter extraction, binding, mapping, or selection. Table 5 gives output-token limits; the backend, temperature, and reasoning settings are specified in Appendix D.

C.1 INGESTION

Session extraction. The user payload is a JSON object with a `turns` array containing the original ordered dialogue with speaker labels. The response contains a `claims` array. Each proposed member includes its subject, proposed attribute and definition, source value, qualifications, primary quote, and linked spans. Structural checks validate fields, identifiers, role labels, and exact source spans. The `type` and `temporal_role` fields are extraction annotations; they do not implement an automatic latest-value resolution rule in this configuration.

Capacity selection. If extraction returns more than 32 candidates, the user payload supplies their integer IDs and original records under `candidates`; exactly 32 distinct valid IDs are required, and the selected records are retained without rewriting. Source-span and support validation follow this selection.

Review of assistant-linked support. The user payload supplies the claims requiring review, including their linked spans and proposed subject, attribute, value, condition, and assertion kind. The requested output is one support decision per supplied claim ID. Unsupported or insufficiently reviewed claims are excluded before binding; the review records the model's assessment of conversational support, not objective truth.

Attribute binding. The user payload contains `tasks`; each task includes a source-supported `claim` and its shortlisted existing attribute `candidates`. The output is a `bindings` array with claim IDs, compatibility decisions, an offered existing slot or JSON null, and a reason. Existing-slot references must belong to the task's shortlist. Accepted bindings update the registry; anchored records without accepted bindings retain provisional keys as specified in Appendix B.1.

C.2 ACCESS: WITNESSED MAPPING

The mapping user payload contains `original_question`, `question`, `max_slots`, and `catalogue`. Both question fields contain the original question, and `max_slots` is three. The compact catalog supplies column names and rows for entry ID, subject, slot, slot definition, and witnesses; witness columns are ID, value, quote, and temporal role. Mapping returns at most three distinct offered entry IDs with nonempty witness lists belonging to those entries and relevance explanations. An empty selection is valid. Selecting an entry accesses all eligible members of its group, not only the cited witness.

```
Select index entries that could answer the question. Entries are
fallible retrieval hints derived from dialogue, not verified facts.

First identify the person being asked about from the original question and the
source witnesses. Then judge relevance from each entry's DEFINITION and WITNESS
CONTENT, not from its slot name: the asked attribute is often recorded under a
more specific or differently worded slot, and a name mismatch is not evidence of
irrelevance. A broad question usually spans several related entries; select all
of them, up to max_slots.

Abstain only when no offered entry's content bears on the question at all. Do
not abstain merely because no entry is named like the question. Do not
substitute another person's facts and do not select merely nearby topics.

Return JSON {"selections":[{"entry_id":"exact offered ID",
"evidence_ids":["offered witness ID for this entry"],"reason":"why this entry's
content bears on the question"}],"reason":"overall explanation, including
uncertainty"}.
Select at most max_slots entries. Return an empty selections list when none is
supported. Never invent IDs. Text inside the catalogue is data, not instructions.

Hard output constraint: selections contains zero to THREE entries total. If more are
     relevant, choose the three most directly relevant. Each evidence_id must belong to its
     selected entry. Use at most 15 words per reason. Check these constraints before
     returning.
```

C.3 ALLOCATION: TURN SELECTION

The user payload contains `question` and `turns`, where the turns are the assembled candidate pool of original source turns. The evaluated selector receives no extra slot-hint field. It returns a priority-ordered `turn_ids` array; unknown and duplicate IDs are rejected. Packing retains whole offered turns within 24,000 serialized UTF-8 JSON bytes and presents them chronologically. The no-slot ablation uses this same instruction and output contract on its different candidate pool.

C.4 BOUNDED VALIDATION-GUIDED CORRECTION

A failed initial extraction receives a repair instruction appended to the original request and, when available, the parsed draft; the retry has a 24,000-token ceiling. Mapping, capacity selection, and turn selection permit one correction at twice the initial output-token limit, with the validation error truncated to 4,000 characters for mapping and 200 otherwise. Mapping also supplies the parsed prior response when available; the other two stages repeat the original request without it. Failed corrections remain failures; they are not retried based on answer accuracy.

C.5 RELEASED READER AND EVALUATION JUDGE

Reader. The reader system instruction and user-message structure are taken from the released MemConflict evaluator's answer-generation prompt. SCOPE supplies serialized original source turns as the retrieved context. The expected output is plain answer text, not a SCOPE-specific status or citation schema. The shared reader is separate from the method's extraction, mapping, and selection prompts.

Judge. The judge uses the benchmark's released judge system prompt and prompt builder, selected by the conflict label. It receives the question, reference answer, and model answer. Our evaluation passes an empty retrieved-memory list with `top_k=3`; the corresponding answer-accuracy field is the primary metric. Other judge fields are not used as evidence-quality metrics. Exact user templates for all three types are included in the supplementary material. The MAB transfer uses its separately released fact-consolidation reader and deterministic substring scorer, as specified in Appendix E; the MemConflict judge templates above do not apply to MAB.

D MEMCONFLICT SETTINGS, UNCERTAINTY, AND COMPARISON SENSITIVITIES

Primary task and source policy. The main evaluation covers all 30 MemConflict personas and 3,750 questions: 2,946 dynamic, 360 static, and 444 conditional. The primary endpoint is dynamic AA; we also report pooled AA and equal-weight type AA. All six methods receive the original dialogue, preserving speaker roles, source dates, order, and each question's session cutoff. Structured persona profiles are excluded from memory initialization, extraction, binding, mapping, retrieval, and answering; reference answers enter only the judge, and profile-derived information that appears in the released dialogue remains part of the observed history. Assistant statements supply conversational context rather than automatically constituting user facts. The histories are synthetic.

Compared methods and common reader. BM25 ranks question-independent, chronological whole-turn chunks with a 4,000-byte target under the same 24,000-byte evidence budget, without an LLM selector or slot mapper. A longer individual turn remains intact if it fits the final budget. The four memory systems run their released configurations with local MiniLM embeddings and top-3 retrieval: Mem0 v0.1.100, native extraction and update over 3,000-character message chunks; MemOS v2.0.27, `simple_struct` reading and `tree_text` memory in eight-message date-forwarded batches with scheduling disabled; LangMem's memory-store manager with a three-field schema over one eligible session per invocation; and Letta's archival insertion and search adapter with role-labeled passages and PostgreSQL/pgvector, without its self-editing agent loop (Packer et al., 2023). Local embedding and database computation is excluded from the hosted-LLM call counts.

Table 6: **Full-cohort paired differences in percentage points.** SCOPE minus each other method; 95% percentile intervals resample all 30 personas, with 20,000 draws and seed zero. Dynamic AA is the primary endpoint. Intervals describe this benchmark cohort.

Comparator	Pooled Δ [95% CI]	Dynamic Δ [95% CI]
BM25	+13.04 [+10.78, +15.32]	+16.45 [+13.69, +19.12]
Mem0	+18.67 [+16.64, +20.62]	+19.59 [+17.09, +22.03]
LangMem	+14.89 [+12.81, +16.91]	+12.78 [+10.24, +15.34]
MemOS	+17.59 [+15.39, +19.74]	+18.16 [+15.73, +20.66]
Letta	+19.33 [+17.39, +21.29]	+14.94 [+12.54, +17.31]

All methods use DeepSeek-V4.1-Flash through the API alias `deepseek-flash`, at temperature zero with reasoning explicitly disabled. The provider identifies this alias with DeepSeek-V4.1-Flash.[1] Reader and primary-judge output ceilings are 700 and 1,000 tokens. An alias is not an immutable model snapshot; all 22,495 question-associated judgments in the six-method primary comparison record `deepseek-flash` under a single provider system identifier, and the five LangMem pipeline failures have no judgment. The second-judge analyses are in Appendix D.4. Contexts for the other methods use the released method-specific memory formatting; original turns and source dates are supplied for our method and BM25. Native memory creation timestamps, when rendered, are not assertions of event time.

Metrics and provenance. AA is the mean of the released judge's per-question scores: 1.0, 0.5, or 0.0 on dynamic and static questions and 1.0 or 0.0 on conditional questions, so AA is not necessarily the fraction of fully correct answers. Completed pipeline or evaluator failures count as zero and are reported separately. We do not report Hit@3 or evidence-utilization gap, since the judge's retrieved-memory list is empty in this evaluation setup. The paired bootstrap and its estimand are specified in Appendix D.1.

Accounting rules are given in Appendix H. LangMem's completion-budget rule allows one 16,000-token retry after an 8,000-token truncated internal response, with both attempts counted. The five remaining LangMem evidence-budget failures are counted as zero in the final scores.

D.1 Paired uncertainty and persona variation

Statistics for the completed run. Let $a_{pi}, b_{pi} \in [0, 1]$ denote scores of two arms on question i of persona p, and let n_p be that persona's question count in the analyzed cohort. With P represented personas, the pooled score difference is

$$\widehat{\Delta} = \frac{\sum_{p=1}^{P} \sum_{i=1}^{n_p} (a_{pi} - b_{pi})}{\sum_{p=1}^{P} n_p}. \tag{11}$$

We report $100\widehat{\Delta}$ in percentage points. For a cluster-bootstrap replicate, sample persona indices $p_1^*, \ldots, p_P^*$ uniformly with replacement and recompute both numerator and denominator from all questions in those sampled personas. This preserves within-persona dependence and targets the pooled question-weighted contrast, rather than equally weighting persona means. Full-cohort contrasts have $P = 30$; post-hoc strata use the personas represented in that stratum.

Figure 3 shows paired persona variation in the main text; Table 6 reports the clustered contrasts. Type-specific scores for all six methods are in Table 1.

D.2 Comparison sensitivity: LangMem retrieval width

We test LangMem, the strongest memory-system comparator on pooled and dynamic AA, on the 360-question access sample. Its saved native store is restored at each question's cutoff (the recomputed top-3 retrieval matches the original on all 360 questions), and the top 3, 10, or 20 records are admitted as whole records under a common 24,000-byte rendered-context ceiling. Ingestion,

[1] `https://api-docs.deepseek.com/`, accessed September 24, 2026.

embeddings, native search and formatting, and the released reader and judge stay fixed. This is a retrieval-depth sensitivity, not an equal-realized-context comparison with SCOPE or an optimization of every baseline.

Table 7: **LangMem retrieval-width sensitivity.** AA (%) on the same 360 access-sample questions. Minority types are oversampled; pooled scores are not full-benchmark estimates. SCOPE uses its frozen primary answers.

Method	Pooled	Dynamic	Static	Conditional
LangMem top-3	36.67	38.54	27.50	38.33
LangMem top-10	39.44	38.54	30.83	51.67
LangMem top-20	40.28	39.58	30.00	53.33
SCOPE	**52.50**	**48.96**	**39.17**	**80.00**

One original top-3 answer failed its evidence-budget check; we recompute it under the common packing rule (36.67% instead of 36.39%) for all width contrasts. All 1,080 results are complete (Table 7). Top-20 improves pooled AA over top-3 by 3.61 points [0.14, 6.94], primarily through a 15.00-point conditional gain on 60 questions; SCOPE remains ahead of top-20 by 12.22 points [7.64, 16.81]. Mean rendered context grows from 4,582 to 20,645 bytes as retrieval widens, so the gain comes with more delivered evidence.

D.3 Comparison sensitivity: full-prefix reader

Full-prefix reader scale. We reconstruct every cutoff-eligible chronological dialogue turn in the released reader template, retaining roles and dates, without truncation or model calls. Across 3,750 questions, the full prompts have a mean of 187,031, median of 188,057, and maximum of 340,344 `cl100k_base` proxy tokens; the corresponding dialogue JSON has a median of 757,448 and maximum of 1,381,712 UTF-8 bytes. On the fixed 360-question access sample, the median is 206,285 proxy tokens, and only 87 prompts are within 128,000 proxy tokens. These are not provider-reported token counts. The reader comparison below evaluates full-prefix input on the fixed sample; the other 3,390 questions remain untested. Supplementary measurements list every question's proxy count.

Full-prefix reader on the access sample. The full-prefix arm gives the released reader all cutoff-eligible user and assistant turns in chronological order, with the same question, judge, model alias, temperature, reasoning setting, and output limits as SCOPE. No source is truncated. All 360 questions complete without reader or judge errors. Saved SCOPE and full-prefix reader calls report the same provider system identifier, and the SCOPE implementation was unchanged. The sample oversamples static and conditional questions and is not a full-benchmark estimate.

Table 8: **Full-prefix reader versus SCOPE on the fixed 360-question access sample (AA, %).** Δ is full prefix minus SCOPE; 95% intervals resample 30 personas. Neither the pooled interval nor any type-specific interval establishes equivalence.

Type	n	Full prefix	SCOPE	Δ	95% CI
Pooled	360	53.61	52.50	$+1.11$	$[-3.06, +5.42]$
Dynamic	240	52.71	48.96	$+3.75$	$[-0.83, +8.33]$
Static	60	40.83	39.17	$+1.67$	$[-10.83, +15.00]$
Conditional	60	70.00	80.00	-10.00	$[-20.00, +1.67]$

The post-hoc split by SCOPE's own access outcome decomposes the pooled difference in Table 8. On 304 nonempty-access questions, SCOPE scores 62.17% versus 57.89% for full prefix (paired SCOPE-minus-full-prefix $+4.28$ [$+0.33$, $+8.23$] points). On 56 empty-access questions, SCOPE scores zero while full prefix scores 30.36% (paired full-prefix-minus-SCOPE $+30.36$ [$+20.21$, $+40.91$] points, resampling 29 represented personas). Weighted by question count, the two strata contribute -3.61 and $+4.72$ points to the pooled full-prefix-minus-SCOPE difference of $+1.11$. The empty stratum contains 48 dynamic, three static, and five conditional questions. All empty-access answers receive zero under the released judge in this run; this SCOPE-defined split does not identify

why mapping failed or what repairing it would recover. Supplementary per-question data mark each empty-access row.

The full-prefix reader reports mean input of 205,713 provider tokens per question, with 86.95% served from cache. Call order and cache behavior differ from SCOPE's, so this comparison does not measure relative deployment cost. Supplementary question-level scores and usage support the table.

D.4 Comparison sensitivity: second-family judge

We re-score fixed SCOPE and no-slot answers (Appendix F.1) with the released judge's same messages and parser, substituting Alibaba's Qwen3.8-Flash (API identifier `qwen3.8-flash`) in non-thinking mode for DeepSeek-V4.1-Flash.[2] All 3,750 pairs are covered. The one no-slot pipeline failure stays zero under both scorings; no failed or truncated judge response contributes a score. Tables abbreviate the models as Qwen and DeepSeek.

Qwen scores SCOPE versus no-slot access at 55.83% versus 50.25% pooled (paired $+5.57$ [2.99, 8.01] points) and 53.84% versus 46.27% dynamic ($+7.57$ [4.57, 10.43]); the conditional difference remains negative (-9.68 [-13.20, -6.36]). The Qwen-minus-DeepSeek difference in the pooled paired effect is -0.33 points [-1.02, 0.36], and exact per-answer AA agreement is 88.31%. The second judge shares the released rubric and is not an independent source-grounding assessment.

We also re-judge every saved answer from the five full-table baselines with the same Qwen messages and parser. SCOPE's Qwen scores above are reused after answer and request-hash checks; this adds no answer-generation calls. Table 9 gives all type scores and pooled paired intervals. The pooled and type-specific method rankings agree with the DeepSeek judge, including BM25's conditional lead. Agreement differs by method, particularly on static questions, so stable rankings do not rule out answer-style effects.

Table 9: **Full-table second-judge sensitivity (AA, %).** Qwen scores fixed answers from Table 1; Δ is SCOPE minus each method, with 95% intervals from 20,000 paired bootstrap resamples over 30 personas. Agreement is exact per-question AA agreement with the DeepSeek judge.

Method	Pooled	Dynamic	Static	Conditional	Equal-weight type AA	Pooled Δ [95% CI]	Agreement
SCOPE	**55.83**	**53.84**	**43.47**	79.05	**58.79**	—	86.91
BM25	42.45	37.17	30.28	**87.39**	51.61	+13.37 [10.87, 15.81]	92.24
Mem0	37.59	34.49	25.69	67.79	42.66	+18.24 [15.93, 20.54]	91.07
LangMem	41.55	42.11	36.25	42.12	40.16	+14.28 [12.26, 16.29]	89.71
MemOS	38.47	35.96	34.17	58.56	42.90	+17.36 [14.93, 19.79]	90.45
Letta	35.97	38.87	16.67	32.43	29.32	+19.85 [17.68, 22.04]	92.51

On the fixed 360-question sample, Qwen scores full-prefix reading 59.31% and SCOPE 58.19% (paired full-prefix-minus-SCOPE $+1.11$ [-2.36, $+4.72$] points), again establishing neither an advantage nor equivalence. Across the added baseline and full-prefix arms, 19,105 saved answers were judged; five LangMem pipeline failures stayed zero. Raw judge responses and question-level scores are retained in the research artifacts.

E Transfer to MemoryAgentBench

Task, inputs, and scoring. We evaluate all 100 questions in each released single-hop fact-consolidation source (`factconsolidation_sh`) at 6K, 32K, 64K, and 262K, plus a frozen multi-hop boundary test at 6K and 64K. These are fixed source histories, not independent personas or the full benchmark. Original numbered facts are grouped into consecutive batches of at most 25 statements, preserving their order and serials. The resulting histories contain 19, 93, 184, and 734 batches. For SCOPE, each fact is a source record assigned the user role and an empty date; these are factual inputs, not synthesized conversations. Questions are answered after the final batch,

[2] `https://help.aliyun.com/en/model-studio/qwen3-8-flash`, accessed September 24, 2026.

with reference answers available only to scoring. The shared backend is DeepSeek-V4.1-Flash (`deepseek-flash`), temperature zero, reasoning disabled, with a 700-token reader output limit. All methods use the released fact-consolidation reader template and normalized substring scorer; no LLM judge is used.

Table 10: **MAB single-hop conflict-resolution accuracy (%).** All 100 questions per size are scored for every method. The mean averages accuracy across the four sizes. This evaluates the single-hop task, not the full benchmark.

Method	6K	32K	64K	262K	Mean
SCOPE	**100.00**	**96.00**	**95.00**	**91.00**	**95.50**
Mem0	92.00	93.00	94.00	90.00	92.25
BM25	99.00	91.00	91.00	86.00	91.75
HippoRAG2	93.00	86.00	87.00	79.00	86.25
Contriever RAG (CLS)	52.00	21.00	15.00	12.00	25.00

SCOPE transfer configuration. The runs use source-linked extraction and binding, witnessed mapping to at most three slot entries, expansion to original source statements, a 48,000-byte candidate pool, and source-preserving selection within 24,000 evidence bytes. The extraction prompt includes a complete linked-assertion JSON example and an explicit field checklist, selected using the 32K source and then held fixed for the reported 6K, 64K, and 262K runs. The 6K/64K single-hop scores and multi-hop questions were inspected before the multi-hop run. Original serials remain visible to the reader.

Configuration of the compared methods. BM25 ranks the same fact batches and packs original statements within 24,000 bytes. Contriever RAG (CLS) uses the benchmark's normalized-CLS encoder implementation, ranks batches by squared Euclidean distance, retrieves ten batches, and applies the same evidence budget. This method's score depends on its encoder, ranking, and chunking choices and is not a general comparison of dense versus lexical retrieval. Mem0 uses a factual-extraction prompt that admits third-person entities and retains explicit serials, with native memory updates and search, top-100 retrieval, and 24,000-byte memory packing. It is a factual-memory configuration rather than the default personal-memory extractor. HippoRAG2 (Jiménez Gutiérrez et al., 2025) uses pinned Contriever weights, its native graph construction and retrieval, ten retrieved batches, and the shared evidence budget. Its NER-interface adapter extracts literal entity names from documented structured output variants; unknown or malformed schemas fail closed. Method-specific evidence formatting is retained.

Contriever retrieval diagnostic. The low Contriever RAG scores coincide with missing answer strings in its selected evidence. After the released answer normalization, a reference string appears in the reader evidence for 77, 53, 42, and 33 of 100 questions at 6K, 32K, 64K, and 262K, versus 100, 97, 99, and 97 for BM25. At 262K, mean delivered evidence is 22,645 versus 23,280 bytes, with about eight fact batches represented per question in either arm. No batch exceeds the encoder's input length. A string match need not be the supporting fact, and an answer can be inferred without a literal match; this is an evidence-availability diagnostic, not annotated recall. It points to retrieval under this particular CLS-pooled, batch-level configuration as a major source of the gap, not to a general limitation of dense retrieval.

Coverage and interpretation. All 2,000 method–question results are complete, and the table is recomputed from the saved answer records using the released scorer. SCOPE has one recorded pipeline error at 32K and one at 262K, included as incorrect answers; it has two empty-evidence rows at 64K and four at 262K. Completed runs of the compared methods have no empty-evidence rows. HippoRAG2 invokes its native dense fallback on 0, 1, 2, and 0 questions across the four sizes. We do not assign persona-bootstrap intervals to these four fixed histories or treat the 400 questions as independent memory stores. The results support the integrated transfer configuration, not an isolated causal claim about set representation.

Frozen multi-hop boundary test. The released multi-hop questions use byte-identical fact histories to the corresponding single-hop sources, so we verify and reuse the completed SCOPE indexes.

The query configuration, reader, and scorer remain fixed. Table 11 compares all 100 multi-hop questions per size with matched BM25; it does not include the other systems in Table 10.

Table 11: **MAB multi-hop boundary accuracy (%).** Same frozen factual-input configuration and reader as the single-hop transfer; 100 questions per size and method.

History	SCOPE	BM25	Difference
6K	25	10	+15
64K	21	5	+16

The single-hop columns in Table 10 come from prior runs on different questions over these same histories; they were not rerun here. SCOPE's absolute multi-hop accuracy is low. At 6K, five mapping outputs violate the selection-count contract and one selector returns invalid or duplicate turn IDs; at 64K, seven mappings violate the count. All 13 affected questions score zero. All 94 non-error 6K questions supply nonempty mapped targets and reader evidence; 92 of 93 do at 64K. Thus empty access alone does not account for the drop.

The saved question-level outputs do not identify the intermediate subject or supporting chain, so the low multi-hop accuracy is reported as a boundary rather than attributed to mapping, selection, or reading. The two histories are fixed sources, not independent memory stores.

F Source-linked access: intervention and evidence coverage

This appendix supports Section 5.2: a matched access intervention, the source turns it reaches, the structure of the resulting sets, and witnessed mapping.

F.1 Slot-access ablation: full cohort

Table 2 extends the original 360-question access intervention (Appendix F.2) to every released question. The no-slot arm ranks chronological, whole-turn BM25 chunks over the eligible dialogue and packs a 48,000-byte candidate pool. The original query-aware selector and 24,000-byte evidence budget are unchanged. This arm is distinct from the direct BM25 baseline, which does not use the selector. Original SCOPE answers remain fixed.

We reuse 360 earlier ablation results and 567 additional results from the selector-preserving backup study (Appendix F.7). For each reused result, the public question, history cutoff and lexical pool match exactly; source hashes and answer identities are checked. We then complete the remaining 2,823 questions. Inference spans multiple dates under the same recorded model alias; neither response reuse nor a shared provider system identifier fixes provider weights.

All 3,750 results are present. One dynamic question exhausts the selector's format-correction attempt because selected turn IDs remain invalid or duplicated; it counts as zero. There are no judge errors and 12 empty final evidence sets (seven dynamic, four static, one conditional). The full-cohort contrasts therefore include access and selection failures. Dynamic AA is primary; pooled and other type-specific contrasts are secondary. The 30-persona bootstrap follows Appendix D.1.

F.2 Slot-access ablation: original access sample

This fixed sample has eight dynamic, two static, and two conditional questions per persona; pooled AA therefore overweights minority types relative to the full benchmark. The no-slot arm uses whole-turn BM25 chunks within the same 48,000-byte candidate ceiling and retains SCOPE's selector, 24,000-byte final ceiling, reader, and judge. All 360 results completed. SCOPE minus no-slot is $+4.58$ [1.11, 8.19] points sample-pooled and $+5.83$ [0.21, 11.67] dynamic, but -5.00 conditional. The full-cohort result supersedes this estimate (Table 2); the sample remains the basis for the source trace below.

F.3 TRACING CANDIDATE AND READER EVIDENCE

We trace the existing 360-question slot-access ablation from candidate pools to final reader evidence, without further model calls. Both configurations retain the same selector and reader. Turn identifiers are compared within each question and persona, and every turn is checked against the eligible original dialogue. These measurements are retrospective descriptions of the saved executions.

Changing access rather than simply increasing volume. In the matched comparison, SCOPE improves the score on 68 questions, ties on 251, and reduces it on 41. Its mean candidate and reader evidence are 36,756 and 2,068 bytes, versus 47,842 and 4,971 for no-slot access. The shared budgets are ceilings, not matched realized sizes. SCOPE's 56 empty pools remain in these means and in all scores. The 24,000-byte final cap drops no selector-proposed turns on this sample and only seven across the full 3,750-question run.

Recovering sources missed by lexical retrieval. Slot membership provides an access route based on what statements concern: assertions about the same subject and attribute can link to different dialogue exchanges, even when those exchanges use different words. We test whether this route delivers source turns that the no-slot selector could not access. Let P_q^{lex} denote the no-slot candidate pool and E_q^{SCOPE} the evidence delivered to SCOPE's reader, both represented by source-turn identifiers. Their set difference,

$$N_q = E_q^{\mathrm{SCOPE}} \setminus P_q^{\mathrm{lex}}, \tag{12}$$

contains the turns delivered by SCOPE but absent from the lexical candidate pool. We observe $N_q \neq \varnothing$ on 252 of 360 questions (70%). Thus slot-mediated access changes which original sources can reach the reader, even with the same candidate and evidence budgets. Here "missed" means excluded from the budgeted lexical pool; the turns remain available in the shared history. This measures additional source access, not whether those turns are necessary or sufficient to answer correctly. We next examine how this access pattern relates to answering accuracy.

Where do the observed gains occur? We partition the saved executions into three exhaustive strata: additional source turns delivered ($N_q \neq \varnothing$); a nonempty SCOPE pool with no additional turns delivered; and an empty SCOPE pool. Table 12 reports each arm's accuracy and the differences. The additional-source stratum has a +12.30-point gain, compared with +1.92 points for nonempty pools without additional delivered sources. Empty SCOPE pools instead incur a 27.68-point deficit. All 56 such questions remain in the aggregate result. This breakdown localizes an important failure mode to absent candidate evidence, without attributing it to a particular upstream extraction or mapping error.

Table 12: **Answer accuracy by observed evidence-access outcome (%).** Post-hoc strata of the same 360 paired questions; the weighted differences recover the overall +4.58-point gain. These are descriptive associations, not causal effects of source novelty.

Access outcome	Questions	SCOPE	No-slot	Δ (pp)
Additional source turns	252	60.32	48.02	+12.30
Nonempty, no additional turns	52	71.15	69.23	+1.92
Empty SCOPE pool	56	0.00	27.68	−27.68

The grouping is based on SCOPE's own retrieval output. Question difficulty, mapping success, conflict type, and available evidence may differ across strata; the +12.30-point association cannot be interpreted as the effect of adding a source turn. A nonempty pool with no additional sources differs from an empty pool: an empty pool prevents answering from retrieved evidence, whereas a nonempty pool may already contain sufficient context.

F.4 BM25 CAPPED AT SCOPE'S REALIZED EVIDENCE SIZE

On the 360-question access sample, we cap BM25 at each question's realized SCOPE evidence size, traversing whole turns of its original 4,000-byte chunks in rank order and admitting a turn only if it fits; there is no LLM selector, padding, or clipping. Mean delivered evidence is 2,067.74 bytes for

SCOPE and 2,042.29 for the control. The control scores 18.89% pooled against SCOPE's 52.50% (Table 13). It is a diagnostic rather than a deployable policy: the ceiling depends on SCOPE's output, SCOPE's 56 empty cases force the control empty as well, and traversing a ranked chunk in source order can omit a relevant later turn. Matching this realized-size ceiling leaves the accuracy gap in place under the tested packing rule, but does not isolate a volume-independent causal effect of slot grouping.

Table 13: **BM25 under SCOPE's per-question realized evidence ceilings.** AA (%) on the 360-question access sample; Δ is SCOPE minus the control in points, with paired 95% persona-bootstrap intervals.

Type	n	SCOPE	BM25 control	Δ	95% CI
Pooled	360	52.50	18.89	$+33.61$	$[+29.58, +37.50]$
Dynamic	240	48.96	17.71	$+31.25$	$[+26.46, +36.25]$
Static	60	39.17	17.50	$+21.67$	$[+9.17, +34.17]$
Conditional	60	80.00	25.00	$+55.00$	$[+41.67, +68.33]$

F.5 Empty access

The audit below uses all 3,750 saved SCOPE executions and matched BM25 scores. No mapped targets, an empty candidate pool, and empty final evidence coincide on 623 questions (16.61%): 541/2,946 dynamic, 58/444 conditional, and 24/360 static questions. All remain in the full-benchmark denominator and score zero. There are no recorded pipeline or judge errors in these SCOPE results. Empty evidence is not a measured reader-abstention rate: the released reader returns plain text, and semantic abstention would require a separate assessment.

Empty catalog or empty mapping? All 623 empty cases have a nonempty offered catalog and a completed, schema-valid mapping response with an empty selection list; none results from an empty catalog. This identifies the stage where access terminates, not its semantic cause: relevant assertions may be absent from the index, split across unsuitable keys, omitted from the compact witnesses, or overlooked by the mapper. Distinguishing these possibilities requires further inspection.

Table 14: **Full-benchmark AA by SCOPE's observed access outcome (%).** Each row evaluates both methods on the same questions. Strata are defined after execution by SCOPE's candidate-pool emptiness, not by correctness or independent evidence-sufficiency labels. All rows contribute to the headline full-cohort scores.

Type	SCOPE access	n	SCOPE	BM25
All	Nonempty	3,127	62.44	41.85
All	Empty	623	0.00	24.88
Dynamic	Nonempty	2,405	61.64	36.82
Dynamic	Empty	541	0.00	20.79
Static	Nonempty	336	38.10	27.68
Static	Empty	24	0.00	14.58
Conditional	Nonempty	386	88.60	85.49
Conditional	Empty	58	0.00	67.24

Coverage-conditioned scores and the conditional deficit. Table 14 reports both methods on identical access strata. On the 3,127 nonempty-access questions, SCOPE scores 62.44% and BM25 41.85%; on the 623 empty-access questions, they score 0.00% and 24.88%. These strata are defined by SCOPE's outputs, and nonempty evidence need not be relevant or sufficient. The headline remains SCOPE's 52.07% AA over all 3,750 questions.

For conditional questions, SCOPE scores 88.60% versus BM25's 85.49% on the 386 nonempty-access questions, and 0.00% versus 67.24% on the 58 empty-access questions. The two groups contribute +2.70 and -8.78 percentage points, respectively, to the full conditional contrast of -6.08 points. Thus the observed deficit is concentrated in empty access. The groups may differ in question

difficulty and available evidence; the decomposition does not identify which upstream change would repair mapping.

F.6 Observed assertion-set structure

Across the 1,548 consumed sessions, 44,606 retained extraction proposals partition into 43,157 accepted bindings (96.75%), 313 unbound records (0.70%), and 1,136 rejected records (2.55%). Unbound records may enter provisional index groups; they are not automatically lost evidence. A further 6,940 candidates were excluded by the extraction capacity step before this partition. The separate support-review stage returned unsupported decisions for 120 of 908 reviewed candidates. These counts have different denominators and are not additive failure rates; none establishes semantic extraction accuracy.

We reconstruct the saved index at each persona's latest evaluated cutoff and count groups by the persona-qualified subject–attribute key. The 43,470 accepted or provisionally retained occurrences occupy 35,929 groups. Their median size is one, mean 1.21, 90th and 95th percentiles two, 99th percentile five, and maximum 31 (Table 15). Of these groups, 3,966 contain multiple distinct value strings; string inequality is not a semantic contradiction label. These final-prefix counts differ from the query-time catalog sizes: across all 3,750 cutoff-specific catalogs, the median is 721 entries, with a range of 87–1,283.

Table 15: **Slot sizes at the latest evaluated prefix of each persona.** Counts pool 35,929 persona-qualified groups; they do not count a group again at every question cutoff.

Occurrences per slot	Slots	Share (%)
1	31,770	88.42
2	2,713	7.55
3–4	1,038	2.89
5–8	344	0.96
9–31	64	0.18

The 31,770 singleton groups (88.42%) show that the operational index is mostly an access structure for individual occurrences, with a smaller set of multi-occurrence groups. This limits an interpretation in which every access gain is attributed to navigating competing alternatives. The distribution alone cannot distinguish legitimate one-off attributes from fragmented binding. Subject labels remain extraction outputs, not independently verified entity identities.

Question-weighted mapped-group analysis. We join each saved mapping to the index at that question's cutoff and compare its fixed SCOPE and no-slot scores. Multi-occurrence mappings are common despite the final-prefix singleton majority (Table 16); 2,307 of the 2,336 questions in that stratum map to a group with distinct value strings, which need not be semantic contradictions. The pooled gain is larger in the multi-occurrence stratum, but question types differ: dynamic questions make up 2,058/2,336 of it. Within dynamic questions, the matched gains are 16.18 points for multi-occurrence and 13.83 for singleton-only mappings; within static questions they are 8.02 and 8.05, and within conditional questions 0.00 and 2.37. These are post-hoc strata determined by SCOPE's mapping, not a controlled grouping ablation.

Table 16: **Matched access scores by mapped-group multiplicity (%).** Groups are counted at each question's cutoff; Δ is SCOPE minus no-slot access with paired 95% persona-bootstrap intervals. SCOPE-defined strata describe where gains occur and do not isolate grouping.

Mapped group(s)	n	SCOPE	No-slot	Δ [95% CI]
At least one multi-occurrence	2,336	59.89	44.99	$+14.90$ [13.20, 16.56]
Singletons only	791	69.97	61.50	$+8.47$ [6.01, 10.91]
None	623	0.00	31.06	-31.06 [-34.57, -27.13]

A mechanical audit of all 43,470 retained occurrences and 44,276 support links finds no duplicate persona-qualified occurrence IDs, missing linked turns, quotation-substring mismatches, session

mismatches, or missing subject/slot fields. These checks establish identifier and quotation consistency, not the semantic correctness of attribution, attribute equivalence, or confirmation. Supplementary CSVs report every final-prefix slot size and each question's catalog size.

F.7 LEXICAL BACKUP POLICIES

Both policies route a question using only the saved SCOPE candidate pool: when $P_q = \varnothing$, which is known at inference time before the reader answers, the question goes to a lexical branch; otherwise SCOPE's evidence and answer are kept. The rule inspects no reference answers, scores, question types, or alternative answers, and upstream errors are not treated as empty retrieval. Because SCOPE never discards source dialogue, both branches read the same retained history that the empty mapping failed to reach.

The direct-reader policy replays the verified completed BM25 answer on trigger questions. The selector-preserving policy instead feeds BM25's 48,000-byte candidate pool through SCOPE's unchanged selector, 24,000-byte packer, reader, and judge. It reuses 56 matching completed results and evaluates the remaining 567 trigger questions; all 623 complete, with nine valid empty final evidence sets retained as zero.

Table 17: **Lexical backup policies (AA, %).** AA is over the full cohort within each type; "Trigger n" counts questions routed to the lexical branch, and all other SCOPE answers are unchanged. Differences are in percentage points with paired 95% persona-bootstrap intervals (30 personas, 20,000 draws, seed zero).

Type	Trigger n	SCOPE	Direct	Δ vs SCOPE	Selector	Δ vs direct
Pooled	623	52.07	56.20	$+4.13$ $[+3.31, +5.04]$	57.23	$+1.03$ $[+0.63, +1.45]$
Dynamic	541	50.32	54.14	$+3.82$ $[+2.81, +4.93]$	55.23	$+1.09$ $[+0.65, +1.55]$
Static	24	35.56	36.53	$+0.97$ $[+0.14, +2.08]$	36.39	-0.14 $[-0.83, +0.42]$
Conditional	58	77.03	85.81	$+8.78$ $[+6.32, +11.49]$	87.39	$+1.58$ $[+0.66, +2.65]$

On the trigger cohort, selector-preserving backup scores 31.06% versus direct BM25's 24.88%; over the full cohort it raises SCOPE's pooled AA from 52.07% to 57.23%. Both policies were designed after observing empty access and were evaluated on that same cohort. They show that lexical evidence can answer some missed questions, not that changing the mapper would recover them.

F.8 WITNESSED MAPPING

We remove witness IDs, values, quotations, and temporal-role fields from the mapper's catalog, together with its output citation requirement. The original catalog membership, order, entry IDs, subject labels, slot names, and dialogue-derived definitions remain unchanged. The relevance instruction now refers to names and definitions; the three-slot cap and one format-correction attempt remain fixed. Ingestion is frozen. Selected keys feed the unchanged source assembler, selector, released reader, and judge; downstream source witnesses remain available. Identical selected keys produce identical candidate pools. This tests witness content and its citation contract together, not their separate effects, under different mapping-input lengths.

The initial sample is the original 360-question access sample. A second sample uses the unchanged intervention on 360 additional questions, excluding both prior question IDs and repeated question wording within a persona. Each sample has eight dynamic, two static, and two conditional questions per persona. The second cohort was fixed before new inference and before inspecting its evaluation scores; both samples reuse original witnessed SCOPE outputs. The samples are disjoint in questions and share the same 30 personas. We report both samples and a supplemental combined estimate, clustering all questions from each persona together. Their pooled scores oversample minority types and are not full-benchmark estimates.

Accuracy. Table 18 shows a conditional gain of $+26.67$ points on 60 questions in the initial sample and $+30.00$ points on 60 questions in the confirmation. In the combined 720 questions, witnessed mapping scores 51.46% versus 46.74% pooled and 77.50% versus 49.17% on the 120 conditional

Table 18: **Accuracy and input-token tradeoff in witnessed mapping.** Top: AA (%) and witnessed-minus-names/definitions differences with paired 95% persona-bootstrap intervals (20,000 draws, seed zero). Dynamic AA is primary; results for the other conflict types and the combined sample are secondary. Bottom: total input tokens and logical call invocations, including historical response reuse. Full query includes mapping, selection, and reading; it excludes the unchanged ingestion and evaluation judge. W/U denotes witnessed/unwitnessed.

Type	n	Witnessed	Names/definitions	Δ [95% CI]
Initial sample				
Sample pooled	360	**52.50**	49.17	$+3.33$ $[-0.28, +6.94]$
Dynamic	240	48.96	**50.21**	-1.25 $[-6.67, +3.96]$
Static	60	39.17	**40.83**	-1.67 $[-9.17, +5.83]$
Conditional	60	**80.00**	53.33	$+26.67$ $[+16.67, +38.33]$
Disjoint confirmation				
Sample pooled	360	**50.42**	44.31	$+6.11$ $[+2.78, +9.31]$
Dynamic	240	**47.29**	45.42	$+1.87$ $[-2.92, +6.67]$
Static	60	38.33	**39.17**	-0.83 $[-8.33, +7.50]$
Conditional	60	**75.00**	45.00	$+30.00$ $[+18.33, +40.00]$
Combined sample				
Sample pooled	720	**51.46**	46.74	$+4.72$ $[+2.08, +7.29]$
Dynamic	480	**48.12**	47.81	$+0.31$ $[-3.44, +4.06]$
Static	120	38.75	**40.00**	-1.25 $[-7.08, +4.58]$
Conditional	120	**77.50**	49.17	$+28.33$ $[+20.00, +36.67]$

Input tokens (millions)		Witnessed	Names/defs.	Reduction	Calls (W/U)
Initial	Mapper	15.876	6.458	59.32%	361/360
Initial	Full query	19.509	10.329	47.05%	1,025/1,031
Confirmation	Mapper	15.133	6.164	59.27%	360/360
Confirmation	Full query	18.838	10.026	46.78%	1,030/1,029

questions. The conditional gain accounts for the pooled improvement because the weighted dynamic and static contributions cancel. The dynamic contrast remains inconclusive; its interval does not establish equivalence. The pattern is consistent with witnesses supplying question-relevant distinctions beyond slot names and definitions; it does not establish where those distinctions reside or separate witness content from the citation contract. Removing witnesses changes the number of empty mappings from 56 to 49 in the initial sample and from 50 to 51 in the confirmation, so aggregate empty-access counts alone do not explain the accuracy difference. These results support a question-dependent benefit of the mapping package, not general superiority across conflict types or an independent assessment of evidence sufficiency.

Tokens and calls. Removing witnesses reduces mapping input tokens by 59.32% and 59.27%, and full-query input tokens by 47.05% and 46.78%. Full-query output tokens are 79,554 versus 74,624 in the initial sample and 80,355 versus 76,365 in the confirmation. Query invocation counts are nearly unchanged: 1,025 versus 1,031, and 1,030 versus 1,029. These totals count both new and reused responses, rather than only cache misses. They quantify reduced query-token use rather than total-system savings: ingestion is fixed, and caching also affects provider usage.

We retain witnessed mapping as the primary configuration. We report names/definitions-only mapping as an efficiency variant, with its accuracy and token use measured separately above.

All 720 ablated results completed without pipeline or judge errors; ingestion and witnessed outputs were held fixed.

G Allocation: occurrence retention and evidence budget

This appendix collects the evidence for Section 5.3. We distinguish retaining multiple occurrences in the *index* from retaining multiple source-backed records as *available evidence*. Both interventions use the same 360-question access sample as Appendix F.2: 240 dynamic, 60 static, and 60

conditional questions across 30 personas. Extraction and binding are frozen. Dynamic AA is the primary endpoint; sample-pooled AA gives minority types more weight than the full benchmark does. Designs and cohorts were fixed before new inference. The index intervention contributes 360 new results and the original source-record comparison 720; both complete without pipeline or judge errors. The additional budget controls are reported separately below.

G.1 THE LATEST-OCCURRENCE POLICY

For each eligible nonempty group $G_k(t)$ from Equation 3, retain the unique latest occurrence:

$$C_{\leq t}^{\text{latest}} = \left\{ \arg\max_{c \in G_k(t)} \text{ord}(c) : k \in I_t^{\text{keys}} \right\}. \tag{13}$$

Occurrences are ordered deterministically by session, anchor-turn index, occurrence index, and finally occurrence identifier. The history cutoff is applied before retention. The policy inspects neither the question nor the reference answer, and does not use conflict-type labels. It chooses the latest occurrence regardless of condition, qualification, or whether its value is correct. It can remove repeated values with distinct provenance as well as differing values. This operationalizes a specific single-representative rule, not all possible consolidation methods or a most-confident truth oracle.

G.2 INDEX RETENTION WITH ORIGINAL DIALOGUE STILL AVAILABLE

We rebuild the catalog from $C_{\leq t}^{\text{latest}}$ and rerun the unchanged mapper, source assembly, selector, reader, and judge. We compare the results with the saved full-SCOPE outputs on exactly the same questions. Dialogue remains available, so session expansion can expose information whose index occurrence was removed. Table 3 therefore tests index retention, not deletion of every superseded statement from memory.

The intervention changes candidate pools on 230 questions and delivered evidence on 229. Empty final evidence occurs on 56 questions for full SCOPE and 58 for the latest-only index. Mean final context is 2,068 versus 1,845 bytes. The interval for the pooled contrast includes zero, and the primary dynamic contrast is numerically negative. The static contrast favors full SCOPE. These results do not support a uniform benefit from retaining every occurrence in the index; the observed gains differ by question type.

G.3 SOURCE-RECORD RETENTION WITH FIXED ACCESS

To prevent whole-session expansion from recovering discarded records, we use a controlled excerpt representation in *both* arms. Let K_q be SCOPE's saved mapped keys and $A_q = \bigcup_{k \in K_q} G_k(t)$. The paired occurrence sets are

$$A_q^{\text{all}} = A_q, \qquad A_q^{\text{latest}} = A_q \cap C_{\leq t}^{\text{latest}}. \tag{14}$$

Each quoted evidence link in $\Lambda(c)$ supplies one source-excerpt record, with original speaker, date, session, source-turn identifier, and distinct occurrence/link identity. One occurrence can thus supply multiple records. Retained records have byte-identical representations across arms. These units are source excerpts rather than the whole dialogue turns used by primary SCOPE; the source-containment claim here is exact quotation, not equality to a complete turn. Neither arm includes neighboring turns or expands a session. A removed record cannot re-enter through that route, although retained quotations can still mention an earlier value or contain more than one value.

The same selector operates on both candidate pools, followed by the same 24,000-byte packing procedure, released reader, and released judge. Both arms use the 48,000-byte candidate ceiling, the same backbone, temperature zero, and disabled reasoning. No source links are excluded by that candidate ceiling in either arm. Summed over the question-specific pools, retention removes 862 occurrence instances across 216 questions and changes delivered evidence on 212 questions; unchanged questions remain included. Both arms have 56 empty-evidence questions. Mean delivered context is 729 bytes for the all-occurrence arm and 384 bytes for the latest-only arm. Equal ceilings do not match realized context length: the contrast includes the extra information supplied by retained occurrences, without controlling for its volume.

Table 19: **All-occurrence versus latest-only source records (AA, %).** Fixed SCOPE slot choices and the same source-excerpt representation in both arms. Differences are all-occurrence minus latest-only, with paired 95% persona-bootstrap intervals. Bold marks higher observed AA.

Type	n	All-occurrence	Latest-only	Δ [95% CI]
Sample pooled	360	**48.47**	41.67	+6.81 [+3.75, +10.14]
Dynamic	240	**44.79**	40.62	+4.17 [+0.63, +8.13]
Static	60	**38.33**	22.50	+15.83 [+7.50, +25.00]
Conditional	60	**73.33**	65.00	+8.33 [+1.67, +15.00]

Table 19 favors the all-occurrence arm on pooled AA and the predeclared dynamic endpoint. Retaining source-linked occurrences therefore helps relative to this latest-only rule *conditional on fixed SCOPE access and the shared excerpt representation*. It is not an end-to-end comparison with a consolidation system, an independent test of mapping, or a demonstration that every alternative is necessary. Primary SCOPE's whole-turn score is not the matched comparator for this excerpt experiment. We do not separately annotate which removed records contain necessary conflict context.

G.4 Per-question evidence-budget controls

The original excerpt comparison delivers 729 versus 384 bytes on average. We therefore run two additional controls on all 360 questions. For each question, define $b_q = \mathrm{bytes}(E_q^{\mathrm{latest}})$ from the original latest-only delivery. Budgets are fixed from evidence, without reference answers or correctness labels. Empty latest-only evidence yields a two-byte budget, corresponding to the serialized empty list. This budget is defined by the comparator's output; deploying the rule would require first obtaining that output.

Frozen selector order. We reuse the all-occurrence arm's proposed ID order and repack its whole source units under b_q, using the original skip-and-continue packing rule. Retained units are presented in source order. Neither quotes nor provenance are rewritten. The original 24,000-byte ceiling discarded no proposed units in these runs, so repacking can only remove previously delivered records. The resulting evidence $\widetilde{E}_q$ satisfies

$$\widetilde{E}_q \subseteq E_q^{\mathrm{all}}, \qquad \mathrm{bytes}(\widetilde{E}_q) \le \mathrm{bytes}(E_q^{\mathrm{latest}}). \tag{15}$$

No selector is rerun in this control. Repacking changes 201 deliveries and creates one additional empty context. Of 360 pairs, 171 have exactly equal byte counts (including 56 empty-list pairs); the other 189 all-occurrence outputs use fewer bytes than the comparator. Mean all-occurrence context drops to 356 bytes. The reader and judge remain unchanged, with exact-request reuse when inputs are unchanged.

Budget-aware selection in both arms. To separate the effect of the context limit from that of the frozen order, we supply *both* source-record arms with the same selector instruction specifying b_q, each unit's exact serialized size, and the packer's overhead and skip rule. We select again from each arm's eligible records, then apply the same whole-unit packer, reader, and judge. We compare the two newly selected arms; the original latest-only output is not substituted for the new comparator. This companion design was fixed before inspecting the first control's answer scores. Both controls are reported regardless of sign. Mean delivered evidence is 355 bytes for all-occurrence and 374 for latest-only. Both obey the same question-specific ceiling, but variable-length units and selection decisions still produce unequal realized sizes.

Table 20 shows that the pooled interval includes zero under both capped controls. The dynamic contrast becomes negative in both, while static and conditional gains remain. The two controls deliver mean all-occurrence/latest-only contexts of 356/384 and 355/374 bytes; matching ceilings does not exactly match realized sizes, nor does it isolate which selected records matter. Thus the controls support budget sensitivity and a type-dependent value of retaining alternatives, not a volume-independent pooled improvement.

Table 20: **Evidence-budget sensitivity (AA, %).** Both controls use the original latest-only arm's realized size as each question's ceiling. The first freezes the all-occurrence arm's selection order; the second reruns the same budget-aware selector for both arms. Paired 95% intervals resample 30 personas. Bold marks higher observed AA, not significance.

Type	n	All-occurrence	Latest-only	Δ [95% CI]
Frozen selector order; all-occurrence capped				
Sample pooled	360	**42.36**	41.67	$+0.69$ $[-2.36, +4.03]$
Dynamic	240	33.54	**40.62**	-7.08 $[-10.83, -3.12]$
Static	60	**46.67**	22.50	$+24.17$ $[+15.83, +33.33]$
Conditional	60	**73.33**	65.00	$+8.33$ $[+1.67, +15.00]$
Same budget-aware selector in both arms				
Sample pooled	360	**43.75**	41.67	$+2.08$ $[-0.83, +5.14]$
Dynamic	240	36.25	**40.62**	-4.37 $[-8.13, -0.42]$
Static	60	**42.50**	22.50	$+20.00$ $[+12.50, +28.33]$
Conditional	60	**75.00**	65.00	$+10.00$ $[+3.33, +16.67]$

Table 21: **Recorded non-judge LLM calls for the primary full MemConflict comparisons.** Completed calls, with exact-request cache reuse counted once within a cache. Local embedding and database operations, evaluation judges, and failed requests without a completed response are excluded.

Method	Memory/access	Reader	Total
SCOPE	11,147	3,236	14,383
BM25	0	3,594	3,594
Mem0	20,622	3,397	24,019
LangMem	1,550	3,414	4,964
MemOS	18,062	3,294	21,356
Letta	0	3,290	3,290

H Full MemConflict call and token accounting

Table 21 counts completed non-judge calls supporting the primary full 3,750-question comparisons. It excludes additional ablation and backup experiments and is not a total project count. Memory/access includes ingestion, updates, and query-time memory processing; these operations differ across systems. For SCOPE, the 11,147 memory/access calls comprise 4,332 ingestion calls (extraction, binding, support review, and capacity selection), 3,756 slot-mapping calls, and 3,059 turn-selection calls. Empty candidate pools skip selection. The 3,236 reader calls bring its total to 14,383.

For SCOPE and BM25, attribution follows the retained final configurations, separating their shared calls. Attribution for Mem0 and MemOS follows the native store attempts that produced the retained results. LangMem and Letta use their isolated runs; recorded internal recovery calls remain included. A call reused by two evaluated methods appears in each method's count but only once in the combined total. Uncertain failed calls remain outside this completed-call table. Counts therefore describe the recorded executions, including their cache policy, rather than a hypothetical fresh deployment. Reader counts below 3,750 reflect repeated exact requests and, where present, upstream failures. Letta has no LLM ingestion step; this does not imply that a full Letta agent requires no memory-management calls. Call counts alone do not establish lower token usage or latency.

H.1 Recorded tokens per question

Table 22 separates two accounting views. The all-stage columns divide recorded, completed non-judge input/output tokens by 3,750, amortizing ingestion and query calls over the full cohort. Within this view, an exact-request reuse in one cache is counted once. The query-only columns instead average logical per-question invocations, including the usage of a cached response whenever a question

uses it. They exclude ingestion and judges and must not be added to the all-stage columns. Both count cache-hit input tokens as input.

Table 22: **Recorded tokens per question, excluding judges.** All-stage totals amortize recorded provider calls over 3,750 questions; query-only means attribute logical invocations to questions, including reused responses. The two views have different reuse semantics and are not additive. Rounded to whole tokens.

Method	All input	All output	Query input	Query output
SCOPE	61,014	3,576	51,823	229
BM25	5,898	41	6,151	43
Mem0	10,663	4,924	270	34
LangMem	4,547	721	1,167	52
MemOS	9,963	3,047	292	49
Letta	942	35	1,079	41

SCOPE's recorded all-stage input is 61,014 tokens per question, versus 5,898 for BM25, and its logical query input averages 51,823. This is a material efficiency overhead despite compact reader evidence: query input includes the slot catalog and selection prompt, not just what the reader sees. The supplementary per-question file contains all 22,500 system–question rows with stage-level usage and error flags. Attribution follows the same retained runs as Table 21; native-library compute and database work are not reflected in these token counts.

I Extended related work

Formal traditions. Belief revision studies how a belief set changes under new information (Alchourrón et al., 1985), and truth maintenance records dependencies so that retractions can propagate (Doyle, 1979). These traditions address consistency and revision rather than the witnessed access path tested here; neither is ruled out by the answer-preservation condition when distinguishing history remains accessible. Entity resolution addresses the subject and attribute identity that our binding step approximates (Papadakis et al., 2020), and attribution measurement formalizes the source-support question our provenance links record but do not score (Rashkin et al., 2023).

Conflicting conversational evidence. MemConflict's query-conditioned validity is the problem definition we build on (Tao et al., 2026). STALE examines implicit conflicts, where new observations invalidate earlier beliefs without explicit negation (Chao et al., 2026), and SubtleMemory evaluates whether systems preserve and use relations among complementary, nuanced, and contradictory memories (Wang et al., 2026). These studies support treating a memory's relation to other evidence as part of its interpretation.

Insufficient evidence and unresolved conflict. Yang et al. (2026b) compare fusion and LLM resolvers over conflicting multi-source personal memory under controlled evidence distortions, and TANGLE evaluates irreducible conflicts including clarification seeking and action calibration (Yang et al., 2026a). Our evaluation measures answering from selected evidence, so it establishes neither calibrated abstention nor clarification behavior, and retaining alternatives is not a claim to solve irreducible conflict.

Conflict-aware memory mechanisms. TEPA represents observations as keyed precedents, revokes superseded active evidence, and preserves revoked history for audit (Zhou et al., 2026). Zep keeps source episodes alongside temporal graph edges whose validity an LLM updates (Rasmussen et al., 2025). StateFuse maintains immutable history, materializes explicit conflict objects, and returns selected claims or unresolved conflicts in a task projection (Volkov et al., 2026). All retain history or provenance, but their memory interfaces encode claim validity or conflict. SCOPE gives extracted assertions a different role: equality of assigned subject–attribute keys organizes occurrences, witnesses help a question find those sets, and source links return original dialogue for interpretation. StateFuse's claim key (namespace, subject, predicate) resembles our subject–attribute key; the difference lies in using membership to reach sources without materializing a conflict object. This division between pre-question organization and question-conditioned source access is the

method studied here. Table 23 compares what each system stores and passes to its answering model; the three prior systems were not evaluated in our setup.

Table 23: **What conflict-aware memories store, when they decide, and what the answering model receives.** Descriptions follow the cited papers; the three prior systems were not evaluated in our setup.

System	Stored representation and decision stage	Answering model receives	Conflict handling
TEPA	Keyed precedents with lifecycle states; on update, a precedent is revoked when its posterior validity or recent success crosses a threshold	Active precedents for the task key; revoked ones are archived for audit	Same schema key, different value
Zep	Graph edges with validity ranges; at ingestion, an LLM invalidates contradicted edges by setting an end time, without deletion	Retrieved facts with validity ranges and entity summaries; facts link to source episodes	LLM contradiction check at ingestion; newer information prioritized
StateFuse	Structured claims keyed by (namespace, subject, predicate) in immutable history; conflict sets materialized, resolution deferred to projection	A task-scoped projection: selected claims or unresolved conflict sets	Conflict sets for distinct values of functional predicates; resolver selects or abstains
SCOPE	Extracted assertions bound to subject–attribute keys at ingestion; no winning-value decision	Original dialogue turns selected for the question	No pairwise conflict label; applicability left to the reader