

Suggested Is Not Superseded: A Pre-Registered Evaluation of a System-One Model as a Memory-Retirement Gate

Xupeng Chen¹ Chenqian Le²

Abstract

Long-term memory for assistants needs a gate that decides when a new utterance retires an old stored claim. Benchmark labels for this step mark a pair as changed or not, so they cannot tell a gate that acts on established change from one that acts on a hint. We write a task contract that separates the two (SUPERSEDED vs. NEEDS-CONFIRMATION vs. COMPATIBLE), label 500 test pairs built from STALE (change pairs plus same-user controls) with three LLM annotators and two human adjudicators (one author and a second annotator; provenance reported), and evaluate Jev, a typed “System One” decision model, under pre-registered hypotheses. Of the 179 test pairs STALE marks as changed, 72 meet the contract’s bar for an established change. Jev never retired a claim on 141 same-conversation non-changes. Its failure is premature retirement: the binary gate retires 39.2% of claims whose change is only suggested. A three-way decision cuts this to 15.5% at a 15.3-point detection cost, which misses our pre-registered 10-point budget, and it lands on the binary gate’s own threshold frontier; these verdicts hold under three readings of the gold labels. On a held-out vault scored against a second annotator’s labels, a cascade that escalated uncertain cases to a reasoning model and looked best on development data was indistinguishable from simply raising the threshold (premature retirement 26.8% vs. 6.5% on dev). Part of the gap comes from the gold labels: relative to that annotator’s reading, LLM-majority gold shares most of its errors with a same-family escalation model, and an identity for binary labels says how much such shared errors hide. The rest is small-sample selection on development data. For builders this means a binary score

should not be wired to delete: route suggested changes to pending evidence, and do not trust a cascade selected on development data of this size.

1. Introduction

An assistant that remembers users must also forget: when a user who said they drive a blue sedan later mentions the silver hatchback, some component has to decide whether the stored claim is retired. Getting this wrong in one direction leaves stale facts in memory; in the other, it deletes true facts on a hint. Recent memory systems and benchmarks (Chhikara et al., 2025; Xu et al., 2025; Wu et al., 2025; Hu et al., 2026) evaluate the first direction well. STALE (Chao et al., 2026) targets it directly: each of its 400 conflict scenarios pairs an early utterance with a later one that, by construction, invalidates it, and it has no non-conflict pairs.

The second direction is harder to measure because the labels do not separate it. Our pilot audit of STALE pairs found that most retrieved “changes” needed an unstated assumption to count as one (a future plan, a different group, two states that can coexist). A gate scored against those labels is rewarded for retiring on suggestion. A new family of fast typed decision models (Almeida, 2026) makes the question practical: a single call returns a score in $[0, 1]$ (a *Noul*) or a distribution over named options (a *Choice*) (TypeSafe AI, 2026), priced by its vendor at \$0.042 per million input tokens (Almeida, 2026), and several groups have begun to use such models as memory controllers, evaluation judges or agent control layers (Jiang et al., 2026a; Li et al., 2026; Wu & Lim, 2026). Jev-Mem (Jiang et al., 2026a), the most complete of these, lets the model govern memory typing, relation construction and retrieval, but by design it keeps every observation rather than making a store-or-discard decision, defers contradiction and obsolescence to a periodic maintenance step that removes nothing, and is scored end to end on question answering. The decision we study, whether a stored claim may be retired, is one that such a system still has to make. We build no memory system and compare none: we evaluate that one decision

¹Dimension Gate ²New York University. Correspondence to: Xupeng Chen <xc1490@nyu.edu>, Chenqian Le <cl6707@nyu.edu>.

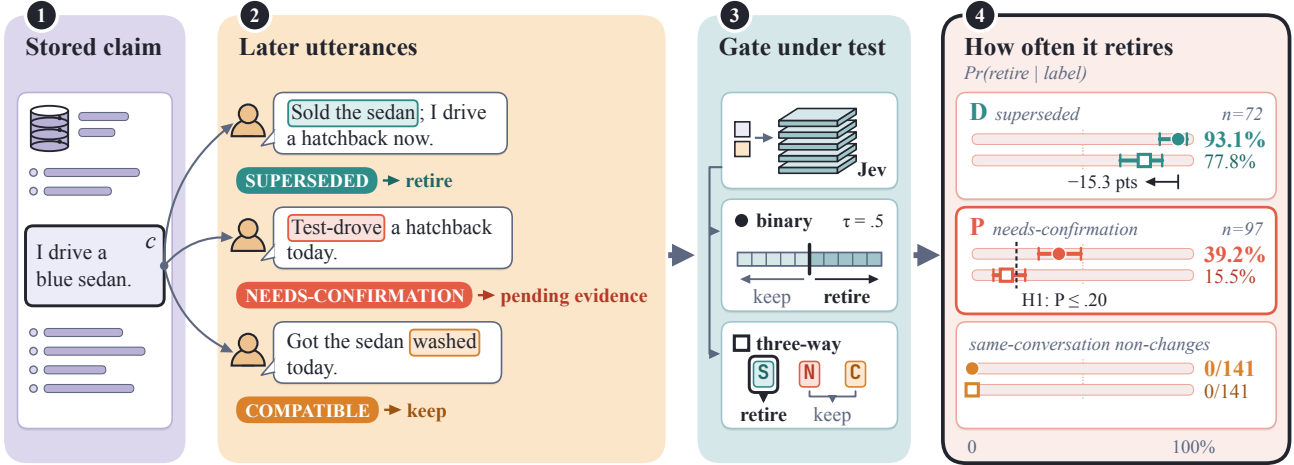


Figure 1. A gate that retired none of 141 same-conversation non-changes still retires stored claims on a hint: as a binary gate it retires 39.2% of suggested changes. (1) A stored claim c . (2) Three later utterances, one per label of the contract (Section 2), with the evidence phrase boxed and the action the contract asks for; the utterances are illustrative, written for this figure. (3) The gate under test: Jev reads c and the utterance, as a binary Noul that retires when its score reaches $\tau=.5$ (S2, filled circles) and as a three-way Choice that retires only on SUPERSEDED (S3, open squares). (4) How often each gate retires test items of a gold label: D on the 72 SUPERSEDED items and P on the 97 NEEDS-CONFIRMATION items, with Wilson intervals as in Table 1 and the pre-registered H1 bar $P \leq .20$ (dashed). The bottom row is the 141 same-conversation controls, a subset of the COMPATIBLE items, of which neither gate retires any (0/141). The pooled rate F over all COMPATIBLE items is in Table 1. The three-way Choice cuts P from 39.2% to 15.5% but loses 15.3 points of detection (93.1% to 77.8%), more than the pre-registered budget allows (H3); Figure 2 shows that it sits on the binary gate’s threshold frontier, a comparison that was not pre-registered. Teal, coral and amber mark the three gold labels.

under a written task contract and pre-registered hypotheses, so that any controller, a maintenance step like Jev-Mem’s included, can be scored on it.

We ask whether such a model can be trusted to retire memories, with the hypotheses fixed in advance; Figure 1 previews the answer. Our contributions:

- A task contract that splits “changed” into established (SUPERSEDED), suggested (NEEDS-CONFIRMATION) and no change (COMPATIBLE), with the actions a memory system should take for each, and a relabelled STALE sample with natural same-conversation controls (Sections 2 and 3).
- A pre-registered evaluation of Jev (Almeida, 2026) against lexical, embedding and reasoning-model baselines. The gate never retires on same-conversation non-changes; it retires on suggested changes, and a three-way decision helps only as far as a threshold shift would (Section 5).
- A held-out replication in which the method selected on development data lost its advantage, and a diagnosis, exploratory because it rests on a second annotator’s labels: gold labels from a majority of LLM annotators share most of their errors with a same-family escalation model, which an exact identity quantifies (Section 6).

2. Task and estimands

A pair consists of a target claim c stated by a user in utterance O_{old} at time t_{old} and a later utterance O_{new} at $t_{\text{new}} \geq t_{\text{old}}$. The label $y \in \{\text{SUPERSEDED}, \text{NEEDS-CONFIRMATION}, \text{COMPATIBLE}\}$ answers: as of t_{new} , what does O_{new} establish about c ? The label is SUPERSEDED only when five conditions hold without any unstated assumption: same subject, same attribute, O_{new} asserts a current or completed state (not a plan, wish or question), same scope, and the new value is incompatible with c . It is NEEDS-CONFIRMATION when O_{new} is evidence that c may have changed but reaching SUPERSEDED needs an assumption, and COMPATIBLE otherwise. Annotators get one test: *if you have to write “assuming...” to reach SUPERSEDED, the label is NEEDS-CONFIRMATION*. A memory system may retire c (and only c) on SUPERSEDED, should attach O_{new} as pending evidence on NEEDS-CONFIRMATION, and should do nothing on COMPATIBLE.

A gate outputs a decision $r \in \{0, 1\}$ (retire or keep). We estimate three conditional rates:

$$\begin{aligned} D &= \Pr(r=1 \mid y=S), & P &= \Pr(r=1 \mid y=N), \\ F &= \Pr(r=1 \mid y=C), \end{aligned} \quad (1)$$

with S, N, C abbreviating the three labels: detection, premature retirement and false invalidation, each with a Wilson 95% interval (Wilson, 1927). The pre-registered hypotheses were fixed before any sample was drawn; the within-

session clause of H1 and the scoring rules were added in dated amendments after sampling, before any human label or system output had been read: **H1** (safety) a gate may retire automatically only if $F \leq .05$, both pooled and on within-session controls, and $P \leq .20$; **H2** (implicit gap) on SUPERSEDED items, Jev’s detection on implicit or numeric changes is at least 20 points below explicit ones and the reasoning model beats it there (McNemar $p < .05$); **H3** (three-way helps) the three-way Choice has lower P than the binary Noul at a detection loss of at most 10 points. Binary gates retire at $\tau = .5$; three-way gates retire only on SUPERSEDED. Strata with $n < 15$ are descriptive. All p -values are exact McNemar tests on paired decisions (McNemar, 1947), unadjusted; the pre-registered tests are H1–H3, and everything else is labelled exploratory.

3. Data and gold labels

Pairs. We drew from STALE’s 400 conflict scenarios (Chao et al., 2026), each an old-state and a new-state utterance, excluding every scenario we had touched in pilot work, and split the rest by scenario into development, test and a sealed vault (49, 179 and 63 scenarios yield pairs). Each scenario gives a *change pair* (STALE’s old and new utterances) and two controls that STALE’s construction implies should not be changes: a *cross-session* neighbour, the user turn between the two sessions most similar to O_{old} under MiniLM (all-MiniLM-L6-v2; Reimers & Gurevych, 2019; Wang et al., 2020), and a *within-session* neighbour, the most similar later turn in the old session itself. The cross-session neighbour turned out to be mostly off-topic (median cosine .29 to the claim), so we added the within-session one (median .39) before any annotation and treat it as the primary false-invalidating control. Test has 500 pairs (179 change, 179 cross-session, 142 within-session); the vault has 176 (63, 63, 50).

Annotation. Three LLM annotators labelled every pair blind to systems and to each other, from the contract text alone: A1, Claude Sonnet 5 through the vendor’s local command-line client; A2, separate Claude subagents; A3, GPT-6 Astra, added because A1 and A2 share a model family (identifiers in Section D). On change pairs A1 agrees poorly with both others (Cohen’s κ (Cohen, 1960) 0.28 with A2, 0.22 with A3) and A2 and A3 agree moderately (0.60); matrices are in Section C, and most disagreement is on the SUPERSEDED/NEEDS-CONFIRMATION boundary that decides retirement. Gold is the three-way majority, with human labels where a human adjudicated. The contract was amended in eight dated addenda, the last three of them after human labels existed; all are released with the contract. One author labelled a first audit batch of 41 items without opening the annotators’ labels (the annotation page logs no reveal; system outputs existed on disk

but had not been summarised). A pre-registered fallback applied: that author overturned the majority on 8/20 items where A3 cast the deciding vote (threshold 15%), so every such test item went to human adjudication (dev items kept majority gold, Section 6). On a random audit of the remaining items the author overturned 5/20. Test gold is 72 SUPERSEDED, 97 NEEDS-CONFIRMATION and 331 COMPATIBLE, with 72 items human-labelled. On the vault every item with any SUPERSEDED vote (37) was labelled by the second annotator described next.

Provenance of the 68 second-annotator labels. The 31 remaining fallback items on test and the 37 vault items were first filled in on the adjudication pages by a coding agent (Codex) running GPT-6 Astra, the model behind annotator A3. Both pages embed the three LLM annotators’ labels behind a reveal control; the agent reported having read them from the pages’ source for the test items and not for the vault items, a self-report we could not verify. We discarded the agent’s labels. One author then relabelled the 68 items, and after that relabel had been compared with the agent’s labels, a second annotator recruited by the authors labelled them on a page that contains no model labels, attesting to working without AI tools or discussion. The rule that gold for these items is the second annotator’s first submission was written before that annotation; after the agreement with the agent’s labels was known, we first set the rule aside and then applied it by the authors’ decision. On the three-way label the second annotator agrees with the agent on 65/68 items and with the author’s relabel on 64/68; requiring the change-expression field to match as well, the figures are 47/68 and 52/68. The author’s relabel agrees with the agent on 67/68 (58/68 with change expression). On the retire-or-keep decision that every number in Section 6 uses, the second annotator and the agent agree on 68/68 (37/37 on the vault) and the author’s relabel on 67/68. The second annotator agrees with A1, A2, A3 and their majority on 23, 44, 54 and 48 of the 68 (the three annotators agree pairwise on 19, 29 and 42; matrices in Section C); the 68 labels were committed over 11.5 minutes, a median of 6 s after each item was first shown. For scale, GPT-6 Astra run on the author’s 41 first-batch items with the same written instructions and no labels matched the author’s three-way label on 26/41 and the retire decision on 29/41. We therefore report the 68 labels as a reading that two people confirmed after a model fill, not as an independent human reading: the test verdicts do not depend on them (Section 5); the vault analysis does, and we label it exploratory.

STALE’s labels. All 179 test change pairs are “changed” in STALE; 72 of them are SUPERSEDED under the contract, and the rest are mostly NEEDS-CONFIRMATION. The split follows STALE’s two conflict types. Of the 89 Type I pairs, which change the same attribute, 59 are SUPERSEDED; of

the 90 Type II pairs, which change a related attribute whose consequence invalidates the old claim, 13 are SUPERSEDED and 66 are NEEDS-CONFIRMATION. The contract’s same-attribute condition excludes Type II propagation by construction, so “most STALE changes are only suggested” is in part a definitional statement about Type II; on Type I it is an empirical one (30 of 89 are not SUPERSEDED). STALE scores answers across three probes; we score one write-time decision on its pairs. We report STALE’s labels separately and do not replace them: they answer a different question (did the topic move?) from ours (may the claim be deleted?).

4. Systems

S0 retires when O_{new} contains a change cue and its word overlap with O_{old} passes a cutoff. **S1** retires when the MiniLM cosine between c and O_{new} is at least .5; both cutoffs were fixed before any output was read. **S2** is Jev (jev-1.13.0) as a binary Noul with the wording of an existing archive-change judge, unchanged. **S3** is a Jev three-way Choice whose options carry the contract’s label names and definitions, developed on dev only. Because Jev’s choices are known to depend on option names (Sun et al., 2026), **S3n** replaces the names with neutral letters and **S3x** gives each definition another label’s name (the SUPERSEDED definition is named COMPATIBLE, and so on). The decision is always read from the definition chosen, not the name. **S4** is Claude Sonnet 5 through the vendor’s local command-line client with the contract text, run on change pairs only. All systems were frozen, and their test outputs stayed sealed until the first scoring. Gold was afterwards revised twice, for the provenance reason in Section 3, by annotators who saw no system output; the first scoring’s test aggregates also informed the method round (Section 6). No system was re-run and no verdict changed (sensitivity in Section 5).

5. Pre-registered results on test

Table 1 and Figure 2a give the pre-registered results.

H1: the gate is safe on non-changes and unsafe on hints.

No Jev variant retired any of the 313 COMPATIBLE neighbour controls, including the 141 same-conversation ones, while the MiniLM baseline retired 21.3% of the within-session ones. Pooled false invalidation is low because most COMPATIBLE items are controls. The 18 change pairs that STALE calls changed but gold calls COMPATIBLE are harder: S2 retires 4 of them, S3 3 and S4 2, too few to separate the systems. The binary Noul fails H1 on premature retirement: it retires 39.2% of claims whose change is only suggested. The three-way Choice (15.5%, interval [9.6, 24.0]) passes on the point estimate, the reading the

contract fixed; the interval includes the .20 bar.

H3: a three-way decision helps less than pre-registered.

S3 cuts premature retirement from 39.2% to 15.5% (McNemar 24:1, $p = 1.5 \times 10^{-6}$) but loses 15.3 points of detection, over the 10-point budget; H3 fails. The two gates rank items equally well (AUROC 0.963 and 0.957), and S3 sits on S2’s threshold sweep: S2 at $\tau=.65$ reaches 79.2%/17.5% against S3’s 77.8%/15.5%. This comparison was not pre-registered. On this data the three-way option acts as a threshold shift, not as extra information.

Sensitivity to the gold reading (exploratory).

With the 31 second-annotator test items scored by the LLM majority instead, S2 retires 39.2% and S3 16.7% of NEEDS-CONFIRMATION items ($n=102$) and S3 loses 15.5 points of detection; with no human labels at all, 41.0%, 17.0% ($n=100$) and 12.7 points. Every H1 and H3 verdict is unchanged under all three readings.

H2: not testable.

Only 11 SUPERSEDED items in the H2 set (human-labelled or unanimous across the three annotators; 9 and 2 of the 11) are implicit or numeric. Descriptively the direction is reversed (S2 90.9%, S4 63.6%). The contract sends most implicit changes to NEEDS-CONFIRMATION by definition, so an implicit-change test needs a different sample.

Option names move the decision, not the ranking.

Neutral names change 3.6% of S3’s decisions and permuted names 10.2%, and detection falls from 77.8% to 51.4%, while AUROC stays within .01 of S3’s. A deployment that reads Jev’s argmax inherits its option-name bias (Sun et al., 2026); the ranking by its probability for SUPERSEDED is nearly unaffected.

Dividing out an option prior does not remove the name effect (exploratory).

For open-weight models, position bias in typed decisions is removed by rotating options and dividing out a prior estimated without labels (Zhang et al., 2026a;b). We applied the same label-free division to Jev’s outputs, estimating each variant’s mean option probabilities on dev and dividing them out on test. The priors barely differ across namings (mean $P(\text{SUPERSEDED})$.14 for S3, .11 for S3x). After division S3x still disagrees with S3 on 11.4% of items (10.2% before; decisions read as the argmax of the returned probabilities), and its detection is 59.7% against S3’s 80.6%. The name effect varies from item to item rather than acting as a fixed per-option offset, so a correction built for position bias does not reach it.

Reasoning-model reference. On change pairs S4 is at parity with S3 (80.6%/14.6%). S4 shares a model family

Table 1. Test results ($n=500$). Rates in % with Wilson 95% intervals; false invalidation pooled over COMPATIBLE items, within-session on its 141 COMPATIBLE controls. AUROC separates SUPERSEDED from the rest. [†]S4 ran on the 179 change pairs only: its premature rate is over the 89 NEEDS-CONFIRMATION change pairs and its false-invalidation rate over the 18 COMPATIBLE change pairs, not the column denominators; on the same 18 items S2 retires 4 and S3 3.

System	Detection $\uparrow$ $n=72$	Premature $\downarrow$ $n=97$	False inv. $\downarrow$ $n=331$	Within-sess.	AUROC
S0 Rule: O_{old} overlap	33.3 [23.5, 44.8]	21.6 [14.6, 30.8]	5.1	8.5	–
S1 MiniLM cosine	19.4 [11.9, 30.0]	5.1 [2.2, 11.5]	9.7	21.3	0.542
S2 Jev Noul, $\tau=.5$	93.1 [84.8, 97.0]	39.2 [30.0, 49.1]	1.2	0.0	0.963
S3 Jev Choice (rubric names)	77.8 [66.9, 85.8]	15.5 [9.6, 24.0]	0.9	0.0	0.957
S3n neutral names	69.4 [58.1, 78.9]	15.5 [9.6, 24.0]	0.9	0.0	0.966
S3x swapped names	51.4 [40.1, 62.6]	6.2 [2.9, 12.8]	0.3	0.0	0.962
S4 Claude Sonnet 5 (change pairs) [†]	80.6 [70.0, 88.0]	14.6 [8.7, 23.4] ($n=89$)	11.1 ($n=18$)	–	–

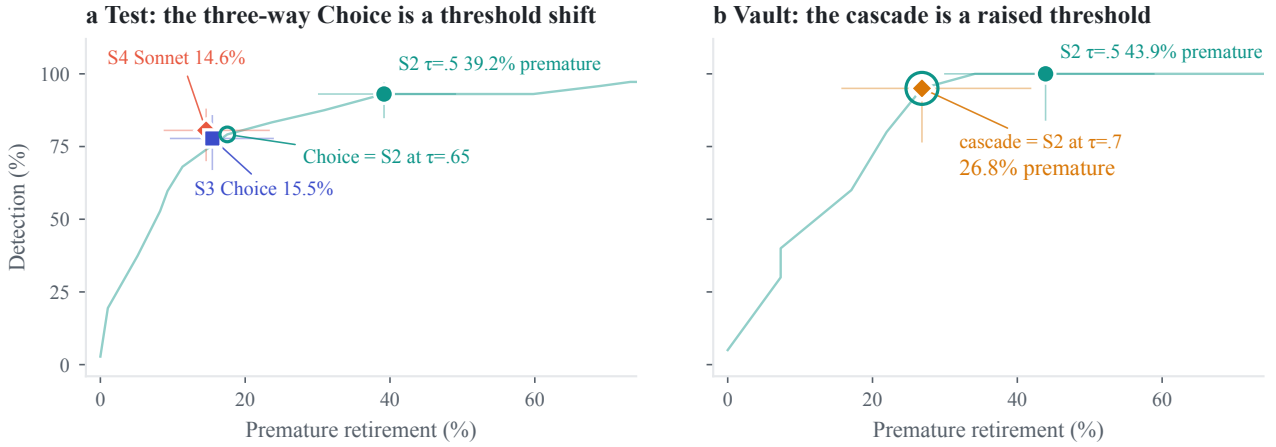


Figure 2. On test the three-way Choice sits on the binary gate’s own threshold frontier, and on the held-out vault the cascade selected on development data coincides with a raised threshold. Detection vs. premature retirement (Wilson 95% cross-hairs); the thin line sweeps S2’s threshold from .05 to .95. (a) Test (72 SUPERSEDED, 97 NEEDS-CONFIRMATION): S2 at $\tau=.5$ 93.1%/39.2%, the three-way Choice S3 77.8%/15.5%, S4 80.6%/14.6%; the ring marks S2 at $\tau=.65$, 79.2%/17.5%, a comparison that was not pre-registered. S0, S1, S3n and S3x are in Table 1. (b) Vault (20 SUPERSEDED, 41 NEEDS-CONFIRMATION, scored against the second-annotator labels, Section 3): S2 at $\tau=.5$ 100.0%/43.9%; the cascade M4 and S2 at $\tau=.7$ (the ring) have identical rates, 95.0%/26.8%, and their decisions differ on 2 SUPERSEDED and 4 NEEDS-CONFIRMATION items, split evenly. Drawn from the per-item system outputs and gold files described in the Reproducibility Statement (Section A).

and definitions with annotators A1 and A2, which favours it.

6. Held-out evaluation and why development overestimated

Method round. With test closed, we developed methods on dev under a written selection rule (highest detection subject to $P \leq .15$ and $F \leq .05$), then ran the winner and a re-thresholded S2 once on the vault. Decomposing the contract’s five conditions into separate Noulis failed on dev (Section D). The winner was **M4**, a cascade: keep if S2 $< .3$, retire if S2 $\geq .9$, otherwise ask S4. Its band, fixed before any reasoning-model output on dev existed, sent 22.2% of vault items to S4. On dev it reached 94.4% detection with 6.5% premature retirement, against 77.8%/9.7% for

S2 at $\tau=.7$ (M5).

The dev advantage did not replicate. On the vault (Table 2, Figure 2b), M4 and M5 both reach 95.0% detection with 26.8% premature retirement, far above M4’s dev estimate and the .15 target. They differ on 1 SUPERSEDED and 2 NEEDS-CONFIRMATION items in each direction. Both improve on S2 at $\tau=.5$ (premature 43.9%; McNemar 7:0, $p = .016$). On this vault (20 SUPERSEDED, 41 NEEDS-CONFIRMATION) the reasoning model buys nothing that raising the threshold does not: identical rates, and 6 discordant items split evenly. A smaller advantage would not be visible here.

Gold type explains part of the gap. Scoring the same vault outputs against LLM-majority gold instead of the

Table 2. Dev (majority gold, selection) and vault (second-annotator labels on all 37 items with any SUPERSEDED vote, exploratory, Section 3). Rates in %, Wilson 95% intervals. Dev gold is the LLM majority; 16 dev items (8 of the 18 SUPERSEDED, 8 of the 31 NEEDS-CONFIRMATION, 15 of the 30 band items) were decided by A3’s vote and never human-checked, because the fallback was scoped to test.

System (selected on dev)	Dev det. / prem. <i>n</i> =18/31	Vault detection <i>n</i> =20	Vault premature <i>n</i> =41	Vault false inv. <i>n</i> =115
S2 Jev Noul, $\tau=.5$	94.4/38.7	100.0 [83.9, 100.0]	43.9 [29.9, 59.0]	0.9
M5 Jev Noul, $\tau=.7$	77.8/9.7	95.0 [76.4, 99.1]	26.8 [15.7, 41.9]	0.9
M4 Noul + Sonnet on [.3, .9)	94.4/6.5	95.0 [76.4, 99.1]	26.8 [15.7, 41.9]	0.9

second-annotator labels gives M4 premature retirement 7/37 (18.9%) instead of 11/41. The rest of the gap is small-sample selection: dev has only 31 NEEDS-CONFIRMATION items and the winner was chosen among 26 (method, threshold) candidates. A selection rule that certifies its target would have refused M4 (exploratory): the one-sided 95% Clopper–Pearson bound (Clopper & Pearson, 1934) on its dev premature retirement (2/31) is 18.9% (by coincidence the same figure as the majority-gold rate above), above the .15 target, and certifying .15 with 2 errors needs at least 40 NEEDS-CONFIRMATION items. AnyJev bounds its early-stopping rule the same way for the same reason (Zhang et al., 2026a). The part due to gold is structural, and the following identity makes it exact. The identity is elementary; its use is to put a number on what a correlated gold hides.

Proposition 1. *Let $S, G, H \in \{0, 1\}$ (retire or keep) be a system’s decision, a gold decision and the true decision on an item. Then*

$$\Pr(S \neq G) = \Pr(S \neq H) + \Pr(G \neq H) - 2\Pr(S \neq H, G \neq H).$$

Hence the error measured against G understates the true error if and only if $\Pr(S \neq H, G \neq H) > \frac{1}{2}\Pr(G \neq H)$, that is, when the system shares more than half of the gold’s errors.

The proof is two lines (Section B); it needs binary decisions, so the three labels are collapsed to retire (SUPERSEDED) or keep. *Instance (exploratory).* Treating the second annotator’s label as H (one reading, Section 3; it coincides with the discarded pre-fill on every retire decision), the 27 vault band items with such a label give $S \neq H$ on 10, $G \neq H$ on 11, and both on the same 7 items, so $S \neq G$ on $10 + 11 - 2 \cdot 7 = 7$. Relative to that reading, majority gold cannot see the 7 errors it shares with the escalation model, and it charges the model with 4 items on which the model agreed with the annotator; the net understatement is 3 errors out of 10. Of the model’s 10 errors, 9 are retirements the annotator would have kept. On all 39 band items the escalation model agrees with the Claude annotator A2 on 35, with the majority on 32, with A3 on 28 and with A1 on 26; with the second annotator it agrees on 17/27 (Figure 3). We did not test why

it tracks A2 so closely. Nor can this data say whether the 7 shared items are errors of the Claude-family models or of a GPT-aligned label; only an annotator independent of both families can.

7. Related work

Memory update and forgetting. Memory layers for agents store extracted facts and revise them on new input. Mem0 (Chhikara et al., 2025) has an LLM choose add, update, delete or no-op for each candidate fact, so its delete branch is a retirement gate, scored only through downstream answers; A-MEM (Xu et al., 2025) links and evolves notes. Benchmarks test whether an assistant answers with the latest state (Wu et al., 2025; Maharana et al., 2024; Hu et al., 2026), and STALE (Chao et al., 2026) isolates the update step. StateMem (Fan et al., 2026) and work on selective forgetting (Li & Li, 2026) and memory-graph evolution (Jiang et al., 2026b) study how state is kept. These score whether the latest value is recovered; we score whether an earlier one was deleted on sufficient evidence, which requires the three-way labels above.

Fast typed decision models. Jev returns a score or a distribution over named options from one call (Almeida, 2026; TypeSafe AI, 2026). Jev-Mem (Jiang et al., 2026a) uses it as the control plane of a memory system without a retirement operation (its maintenance step flags contradiction and obsolescence but removes nothing), which is why it is not a baseline here; others use it as an evaluation judge (Li et al., 2026) and for selective control of agents (Wu & Lim, 2026); Ibrahim & Zaki (2026) evaluate decision models as text annotators, and Sun et al. (2026) show that its choices depend on option names. AnyJev (Zhang et al., 2026a;b) brings the interface to open-weight models by reading option-token logits, removes position bias with option rotation and a label-free prior, and fits per-question linear heads on intermediate layers; its accuracy is reported on the public JevBench subset against that benchmark’s released expected answers. We replicate the name effect on a new task at a smaller magnitude, show that the ranking by probability is nearly unaffected, and show that the prior division that fixes position bias does not fix it.

Evaluation with LLM labels. Using LLMs as annotators (Gilardi et al., 2023) or judges (Zheng et al., 2023) is routine, and so are two warnings: judges favour outputs of their own family (Panickssery et al., 2024), and when the judge is no more accurate than the model it scores, debiasing it with a few gold labels can at best halve the gold labels an evaluation needs (Dorner et al., 2025). Our contribution here is narrow: a pre-registered fallback that measured the problem (8/20 overturns) and forced human labels where it mattered, and an exact decomposition of how much a same-family gold hides.

8. Limitations

Each human-labelled item has one adjudicator (one author for the first batch, the second annotator for the other 68), so the human labels are one reading, not a consensus; their provenance is given in Section 3. On the vault every human label is the second annotator’s, so the direction of the shared-error finding in Section 6 is conditional on that reading, and we label it exploratory; an annotator independent of both model families would settle it. Gold on the 428 test items not sent to the human is the LLM majority, and the random audit overturned 5/20 of those; the controls, on which no annotator ever voted SUPERSEDED, are the least affected. Instructions to the human were clarified between two batches (stated before the second batch), so human-label statistics should be read per batch. Results with the labels the human revised after discussing batch-one items with an assistant leave every verdict unchanged, and gold uses the first-submission labels throughout. H2 had 11 items. The three-way Choice, the only gate that met H1 on test, was not run on the vault; its test premature rate of 15.5% is therefore unreplicated, and the vault’s shift for S2 at $\tau=.7$ (11.3% on test, 26.8% there) suggests it would not hold. That run is pending, and until it and the independent vault labels are in, the vault comparison is exploratory. All results are for one decision model at one pinned version (jev-1.13.0) on one English benchmark (STALE); a later version is a different instrument, and whether the name effect or the threshold-frontier finding holds for other typed decision models or other update benchmarks is untested. The hosted service is also close to, not exactly, deterministic: identical requests reproduce a score only approximately, so a small share of decisions near a threshold can change between calls, and every Jev decision reported here comes from one request that was not re-sent. The vault is spent and no unused STALE scenarios remain, so the observation that GPT-6 Astra agreed most with the vault second-annotator labels (30/37), labels that match the Astra-driven agent’s fill on 37/37 retire decisions, needs new data. We did not analyse cost or latency, which the contract listed, nor what the gate changes end to end inside a memory system such as Jev-Mem or Mem0.

9. Conclusion

Scored against a contract that separates established from suggested change, a System-One gate is safe on same-conversation non-changes and unsafe on hints: the binary score retires 39.2% of suggested changes, and the three-way decision that cuts this to 15.5% is a threshold shift, not new information. A reasoning-model cascade that looked best on 31 development NEEDS-CONFIRMATION items did not beat the raised threshold on a held-out vault, and part of its apparent advantage was gold labels that share errors with it (an exploratory reading, Section 6). Three practices follow. Memory systems should treat a positive gate decision as evidence to attach, not as permission to delete, unless the decision model has been scored on suggested changes. Benchmarks of memory update should label established and suggested change separately; STALE’s “changed” covers both. Evaluations that use LLM-majority gold should report, per Proposition 1, how many of the system’s errors the gold can see.

Impact Statement

This paper evaluates whether a fast decision model can be trusted to retire stored facts about a user, a step that any assistant with long-term memory must perform. The risk it measures is concrete: a gate that retires on suggested rather than established change deletes true facts on a hint, and benchmark labels that do not separate the two reward that behaviour. The conversational pairs come from a public benchmark; the only human participants were one author and one annotator recruited by the authors, who labelled text pairs from that benchmark. This paper presents work whose goal is to advance the field of Machine Learning; beyond the risk described above, we see no societal consequence that must be specifically highlighted.

Competing interests. The authors have no affiliation with TypeSafe AI and received no funding from it; Jev was accessed through its public API.

References

- Almeida, D. Introducing System One Models & Jev. TypeSafe AI Blog, <https://typesafe.ai/blog/introducing-system-one-models-and-jev>, 2026. 15 September 2026.
- Chao, H., Bai, Y., Sheng, R., Li, T., and Sun, Y. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.
- Chhikara, P., Khant, D., Aryan, S., Singh, T., and Yadav, D. Mem0: Building production-ready AI

- agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.
- Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.
- Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.
- Dorner, F. E., Nastl, V. Y., and Hardt, M. Limits to scalable evaluation at the frontier: LLM as judge won’t beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025.
- Fan, X., Liu, M., Yang, R., Ouyang, S., and Han, J. Can agent memory systems track evolving state? *arXiv preprint arXiv:2608.19652*, 2026.
- Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. doi: 10.1073/pnas.2305016120.
- Hu, Y., Wang, Y., and McAuley, J. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations (ICLR)*, 2026.
- Ibrahim, H. and Zaki, Y. Evaluating Decision Models for Text Annotation in Computational Social Science. *arXiv:2609.24574*, 2026.
- Jiang, D., Li, Y., and Li, B. Jev-Mem: System-One-Controlled Agentic Memory for Efficient AI Agents. *arXiv:2609.23986*, 2026a.
- Jiang, D., Li, Y., Li, G., Li, Q., and Li, B. HAGE: Harnessing agentic memory via RL-driven weighted graph evolution. *arXiv preprint arXiv:2605.09942*, 2026b.
- Li, Y. and Li, Y. What should an agent forget? separating what is stored from what is used. *arXiv preprint arXiv:2609.10263*, 2026.
- Li, Y., Miao, Y., Krishnan, R., and Padman, R. JEV-as-a-judge: Accept when confident, escalate when unsure. *arXiv preprint arXiv:2609.26550*, 2026.
- Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.
- McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.
- Panickssery, A., Bowman, S. R., and Feng, S. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3980–3990, 2019. doi: 10.18653/v1/D19-1410.
- Sun, Y., Xu, J., Shi, J., and Yang, Z. Type-Safe Is Not Error-Free: A Constrained Decision Head Follows the Option Name, Not the Rubric Bound to It. *arXiv:2609.26758*, 2026.
- TypeSafe AI. TypeSafe developer documentation: Primitives (Choice, Noul). <https://docs.typesafe.ai>, 2026. Accessed 2026-10-04.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.
- Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025.
- Wu, T. and Lim, W. Y. B. REFLEX with Jev for efficient selective control in LLM agents. *arXiv preprint arXiv:2609.26532*, 2026.
- Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., and Zhang, Y. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- Zhang, J., Yang, T., Shi, Y., Chen, E., Nie, Z., Wan, K., Hong, L., Liu, N., and Wu, L. AnyJev technical report. *arXiv preprint arXiv:2610.00831*, 2026a.
- Zhang, J., Yang, T., Shi, Y., and Wu, L. AnyJev: Turn any LLM into a Jev-style decision model. <https://github>

[ub.com/nokia-applied-research/AnyJev](https://nokia-applied-research.github.io/AnyJev/),
2026b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z.,
Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H.,
Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge
with MT-Bench and Chatbot Arena. In *Advances in Neu-
ral Information Processing Systems (NeurIPS), Datasets
and Benchmarks Track*, 2023.

A. Reproducibility Statement and Use of AI Assistants

Reproducibility statement. Every number, table and figure in this paper is generated by scripts from the saved annotator and system outputs; a manifest of those input files (sizes and SHA-256 digests) and a claims checker that re-derives each number from the per-item outputs will be released with the code. The task contract and its dated amendments, the pre-registered hypotheses and selection rule, the item identifiers into STALE with every annotator’s labels, the per-item system outputs for test and vault, the gold files (majority votes and every human export) and the code will be released upon publication; the two agent pre-fill exports, which the code repository alone cannot regenerate, ship with it.

Use of AI assistants. AI assistants helped write some of the experiment code and check its correctness, helped collect public data, helped check the correctness of citations, and helped polish the paper. Proposition 1 and its proof were drafted by an AI assistant and checked by the authors and by a separate AI assistant. LLMs also served as annotators, and Section 3 records how the 68 second-annotator labels came about.

B. Proof of Proposition 1

For binary values, $S \neq G$ holds exactly when one of S, G differs from H and the other does not, so $1[S \neq G] = 1[S \neq H] + 1[G \neq H] - 2 \cdot 1[S \neq H] \cdot 1[G \neq H]$ pointwise: the right side is 0 when neither or both differ from H and 1 when exactly one does. Taking expectations gives the identity. Subtracting $\Pr(S \neq H)$, the measured error is smaller than the true error iff $\Pr(G \neq H) - 2 \Pr(S \neq H, G \neq H) < 0$. $\square$

C. Annotator agreement

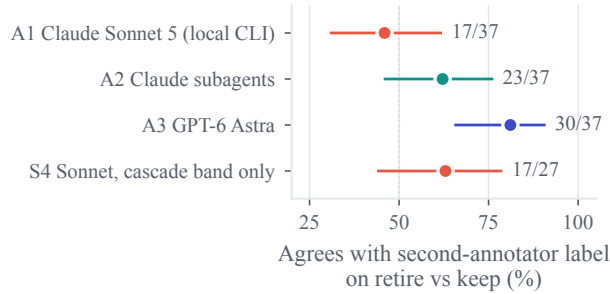


Figure 3. Agreement with the second annotator’s label (exploratory) (Section 3) on retire vs. keep, vault items with any SUPERSEDED vote. Top three rows: annotators on all 37 items; bottom row: S4 on the 27 of them in M4’s band. Wilson 95% intervals. Drawn from the per-item system outputs and gold files described in the Reproducibility Statement (Section A).

Table 3. Confusion matrices on dev and test change pairs ($n=228$); rows are the first annotator.

row \ col	A1-A2 ($\kappa=0.28$, $n=228$)			A1-A3 ($\kappa=0.22$, $n=228$)			A2-A3 ($\kappa=0.60$, $n=228$)		
	SUP	NC	COMP	SUP	NC	COMP	SUP	NC	COMP
SUP	56	17	0	40	33	0	71	35	0
NC	50	77	8	35	86	14	5	96	10
COMP	0	17	3	1	13	6	0	1	10

Table 4. The second annotator’s 68 labels (rows) against A1, A2, A3 and their majority (columns).

second annotator \ col	A1			A2			A3			majority		
	SUP	NC	COMP	SUP	NC	COMP	SUP	NC	COMP	SUP	NC	COMP
SUP	12	25	0	36	1	0	30	7	0	31	6	0
NC	15	11	0	18	8	0	3	23	0	9	17	0
COMP	3	2	0	3	2	0	0	4	1	1	4	0

Detection by STALE conflict type (pre-registered stratum; Type II is descriptive, $n=13 < 15$): S2 96.6% on the 59 Type I SUPERSEDED items and 76.9% on the 13 Type II ones; S3 81.4% and 61.5%.

D. Method-development round and model identifiers

Candidates, written before running: M1, five Nouns, one per contract condition, retire iff all exceed τ ; M2, S2 AND M1; M3, mean of the Jev probabilities; M5, S2 re-thresholded; $\tau \in \{.5, .6, .7, .8\}$. M1 retired almost nothing on dev because the scope question included “time frame”, which every change violates. Post-hoc variants using only the attribute and incompatibility questions (M1b, M2b, M6) did no better than M5. The cascade M4 and its band were written down before any reasoning-model output on dev existed. All dev numbers use majority gold; the 26 candidates and the selection rule are replayed from the per-item outputs by the released code.

Model identifiers. A1 and S4: Claude Sonnet 5 (`claude-sonnet-5`) through the vendor’s local command-line client, one pair per request. A2: separate Claude subagents of the same family. A3: `gpt-6-astra` through its Responses API at reasoning effort *medium*; the coding agent that pre-filled the 68 items ran the same model at effort *high*. S2, S3, S3n, S3x: `jev-1.13.0`. Dev and test were annotated and run on 23–24 September 2026, the vault on 27 September; the second annotator labelled on 28 September.