Does Synthetic Data Generation Strategy Affect Agent Memory Evaluation?

Sebastian-Antonio Toma*
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Sebastian.Toma@campus.utcluj.ro

Oliviu Matei
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

Abstract

Synthetic benchmarks are increasingly used to evaluate AI agent memory systems, yet while recent work has examined bias in synthetic test collections for information retrieval and RAG evaluation, we find none examining whether document generation strategy affects intrinsic evaluation of agent memory systems, where generated artifacts must encode temporal fact evolution across multi-document sequences. We ablate over 11 generation strategies spanning 5 axes (verification strictness, generator model size, style grounding, model diversity, and document realism), holding ground truth, scoring, and the system under test constant. Results show that generation strategy affects absolute scores, with up to 8.1 percentage points difference in fact accuracy across strategies on the same ground truth. Multi-model rotation nearly doubles corpus vocabulary diversity over single-model generation at moderate additional cost, reducing the stylistic homogeneity that can inflate scores. We provide actionable guidance for practitioners building synthetic benchmarks for agent memory evaluation.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**; • **Computing methodologies** → *Natural language processing*.

Keywords

agent memory, evaluation, synthetic data, benchmark, generation strategy

ACM Reference Format:
Sebastian-Antonio Toma and Oliviu Matei. 2026. Does Synthetic Data Generation Strategy Affect Agent Memory Evaluation?. In *The Eighth International Conference on Distributed Artificial Intelligence (DAI '26), November 29–December 2, 2026, Hong Kong, China.* ACM, New York, NY, USA, 11 pages.

1 Introduction

As AI agent memory systems move from research prototypes to production deployment [3, 9, 27], evaluation methodology must keep pace. Synthetic benchmarks, where documents are generated from known ground truth, have emerged as the dominant evaluation paradigm, avoiding the expense and subjectivity of human annotation [17]. Yet a critical assumption remains open: does the *strategy* used to generate benchmark documents affect evaluation conclusions?

Recent work has begun to address this question in adjacent domains. Rahmani et al. [16] show that LLM-generated synthetic test collections introduce measurable bias in information retrieval evaluation, though the effect on relative system rankings may be limited. Van Elburg et al. [22] demonstrate that synthetic RAG benchmarks reliably rank systems by retriever configuration but fail to produce consistent rankings when comparing generator architectures. Maheshwari et al. [11] find that smaller LLMs exhibit biases toward their own generated data, recommending multi-model generation.

Agent memory evaluation presents a distinct challenge. Unlike static retrieval benchmarks, agent memory benchmarks encode *temporal fact evolution*: a fact must be created, reinforced, updated, or superseded [20] across a sequence of multi-author documents. Each document must contain new information, testing whether the system retains prior knowledge across ingestions. The generation strategy must therefore produce documents that reliably encode specific claims at controlled difficulty levels, a constraint absent from query-based IR benchmarks.

We hypothesize that generation strategy introduces systematic bias into agent memory evaluation through three mechanisms: (1) model-specific linguistic behaviors that make extraction easier or harder, (2) style homogeneity that fails to test robustness to natural variation, and (3) document realism that affects how well benchmark performance predicts real-world performance.

To test this, we ablate over 11 generation strategies using the FEB-30 benchmark [19] (30 fact-evolution trajectories spanning 6 diagnostic categories and 3 difficulty levels). All variables except the generated documents are held constant: same trajectories, same ground truth claims, same scoring framework, same system under test. We measure whether absolute scores and difficulty calibration change across strategies, and characterize the generated documents and their cost.

Our contributions are:

(1) An ablation study over 11 generation strategies spanning 5 axes (verification strictness, generator model size, style grounding, model diversity, and document realism), showing that generation strategy affects agent memory evaluation scores.
(2) Identification of which strategy components most affect scores, across-run stability, and document-level characteristics.
(3) A cost-benefit analysis showing which components justify their additional cost for benchmark construction.
(4) Actionable guidelines for practitioners on generation strategy selection for agent memory benchmarks.

*Also with Amazon Web Services.

2 Related Work

Bias in Synthetic Evaluation Data. The reliability of LLM-generated evaluation data is an active area of study. Rahmani et al. [16] conduct a systematic investigation of bias in synthetic test collections for IR, showing that LLM-generated queries and relevance labels introduce measurable distortions in absolute system scores, though relative rankings may be more robust. Van Elburg et al. [22] extend this to RAG evaluation, finding that synthetic benchmarks produce reliable retriever rankings but inconsistent generator rankings, suggesting that generation strategy interacts with the evaluation target. Maheshwari et al. [11] demonstrate self-preference bias: smaller models score higher on benchmarks generated by themselves, while larger models are less affected. Since our pipeline uses LLMs for both verification and scoring, it is also subject to the broader LLM-as-judge biases documented in the literature, including self-preference [15] and positional or verbosity effects [26]; we mitigate the former by keeping the generator pool disjoint from the extraction model.

Beyond evaluation, synthetic data diversity has been shown to matter for model training. Multi-source synthetic data mitigates distribution collapse in fine-tuning [18], and single-model synthetic training data reinforces model-specific blind spots [24]. These findings suggest an analogous risk for evaluation: single-model benchmark generation may produce artifacts with systematic properties that advantage or disadvantage specific system architectures.

Agent Memory Benchmarks. Agent memory evaluation has advanced rapidly. Widely used conversational-memory benchmarks such as LoCoMo [10] and LongMemEval [23] rely on synthetic or semi-synthetic generation and are a natural audience for this study. MemoryAgentBench [5] defines four memory competencies evaluated through QA interactions. Memora [21] introduces freshness-aware metrics decomposing performance into temporal components. MEME [7] introduces controlled state trajectories with multi-entity evolution. AMA-Bench [25] evaluates long-horizon memory for agentic applications. The Anatomy survey [6] catalogs architectures and identifies benchmark saturation concerns.

All these benchmarks use synthetic or semi-synthetic data generation, yet largely leave open whether their generation methodology affects the conclusions drawn. A recent survey documents the gap between response-quality and knowledge-state correctness in agent memory evaluation [20], motivating the intrinsic scoring of FEB-30. FEB-30 [19] adopts paradigm inversion with multi-model rotation and 3-judge verification but presents it as a design choice without ablating alternatives.

Paradigm Inversion. RIKER [17] introduced generating evaluation corpora from known ground truth for knowledge retrieval evaluation, avoiding human annotation. FEB-30 extended this to temporal fact evolution with multi-document sequences. Our work treats the generation pipeline itself as the experimental variable, examining which paradigm inversion choices affect downstream evaluation validity.

3 Methodology

We compare 11 document generation strategies applied to the same benchmark ground truth. The experimental design isolates generation strategy as the sole independent variable.

3.1 Benchmark: FEB-30

We use FEB-30 [19], a benchmark of 30 fact-evolution trajectories organized along two axes: 6 diagnostic categories (LLM compensation, retrieval bottleneck, classification accuracy, synthesis quality, temporal correctness, multi-author attribution) and 3 difficulty levels (L1: claims nearly explicit; L2: claims embedded in context; L3: claims require inference). Each trajectory defines a sequence of states with atomic ground-truth claims and expected memory operations (create, update, reinforce, supersede).

3.2 Generation Strategies

We compare 11 generation strategies organized as a main ablation ladder (S0–S6) and micro-ablations that isolate individual factors (Table 1). The generation prompt is identical across all strategies unless noted; only the strategy-level parameters change. Each step from S1 to S6 adds exactly one component, enabling causal attribution of score differences. The ladder builds up to the FEB-30 configuration and beyond:

S1 (Single model, no style). A single high-capability model produces all documents with 3-judge unanimous verification. No style references are provided. This represents the cheapest verified approach, the default when practitioners build quick evaluation sets.

S2 (+ Style grounding). The same single model generates documents but receives a style reference, a real document from the AMI Meeting Corpus [2] or Enron Email Dataset [8] matched by document type (meeting notes, email, architecture review, status update). The prompt instructs the model to mimic the tone, vocabulary, and structure of the reference.

S3 (+ Multi-model rotation). Six generation models from five providers are rotated per state index within each trajectory, simulating different human authors with different writing styles. Style references are provided as in S2. No single model generates more than 30% of total documents. Multi-model rotation is a component of FEB-30's generation pipeline.

S4 (+ Hybrid injection). For L3 (hardest) trajectories, target claims are injected into real human documents from unrelated domains rather than generating new documents. The model weaves claims as brief asides into an authentic host document (AMI meeting transcript or Enron email about an unrelated topic), producing text that is predominantly authentic human-written content. L1/L2 trajectories use the same generation procedure as S3 (identical model assignment, prompt, and noise count), but are generated independently; because generation is stochastic (sampling temperature and randomized noise selection), S4's L1/L2 documents differ from S3's. Differences between S3 and S4 at L1/L2 therefore reflect generation variance rather than a strategy effect, and the hybrid-injection contrast between them is meaningful only at the aggregate level and at L3, where the strategy actually differs. Hybrid injection is the L3 generation step in FEB-30's pipeline.

ID	Description	Generator	Style Ref	Noise	Judges	Hybrid
Verification ablation (pre-baseline)						
S0	No verification	1 model	×	✓	0	×
S0-1judge	Single-judge verification	1 model	×	✓	1	×
Main ablation ladder						
S1	Single model baseline	1 model	×	✓	3	×
S2	+ Style grounding	1 model	✓	✓	3	×
S3	+ Multi-model rotation	6 models	✓	✓	3	×
S4	+ Hybrid injection	6 models	✓	✓	3	✓
S5	+ Human review	6 models	✓	✓	3	✓
S6	+ Sequential context	6 models	✓	✓	3	✓
Micro-ablations						
S1-lite	Smaller generator	1 model (small)	×	✓	3	×
S1-nonoise	No distractor noise	1 model	×	×	3	×
S4-persona	Named author personas	6 models	✓	✓	3	✓

Table 1: Generation strategies. All strategies use the same 30 trajectories, identical ground truth, and identical scoring. Each main ladder step adds exactly one component to the previous step. Micro-ablations branch off their respective baselines to isolate individual factors.

S5 (+ Human review). A human reviewer inspects all generated documents, corrects factual inaccuracies (e.g., forward leakage, terminology drift), removes residual scaffolding, and ensures natural flow. This represents the gold-standard pipeline where automated generation is followed by manual quality assurance. Documents are the S4 output with human post-processing applied (the corpus released in FEB-30 [19]).

S6 (+ Sequential context). When generating the document for state i, the generator receives the previously generated documents for states $0 \ldots i{-}1$ of the same trajectory as additional context. This enables temporal references ("as discussed last week…") and cross-document consistency, producing coherent multi-document narratives rather than independently generated artifacts. S6 builds on S4 rather than the human-reviewed S5.

S0 and S0-1judge test whether the quality assurance step itself affects evaluation outcomes:

S0 (No verification). Documents are accepted on first generation without any judge verification. This is the fastest and cheapest pipeline: generate once, use immediately.

S0-1judge (Single judge). A single judge model must confirm that all target claims are inferable. Documents failing are regenerated up to 3 times. This tests whether the 3-judge unanimous consensus (used in S1–S4) provides meaningful additional quality over a single judge.

Three additional strategies isolate specific factors:

S1-lite (Smaller generator). Identical to S1 but uses a smaller, cheaper model from the same family. This tests whether generator model size, and the associated capability differences, affects evaluation outcomes. This connects to findings by Maheshwari et al. [11] that smaller models exhibit stronger self-biases.

S1-nonoise (No distractor noise). Identical to S1 but removes all distractor noise claims from L2/L3 documents. Documents contain only target claims in surrounding context. This tests whether noise calibration (which controls extraction difficulty) affects evaluation scores independently of other strategy choices.

S4-persona (Named author personas). Identical to S4 but each model in the rotation receives a named author persona: a consistent name, role, writing style description, and vocabulary preference (e.g., "You are Sarah, a detail-oriented Solutions Architect who writes concise technical emails with bullet points"). This tests whether explicit author simulation on top of natural model diversity produces more realistic variation and affects evaluation outcomes.

3.3 Controlled Variables

Across all strategies, the following are held constant:

- **Ground truth:** Same 30 trajectories, same claims per state, same expected operations.
- **Generation prompt:** Identical prompt template including identifier-preservation constraint, document type, author role, noise injection (where applicable), and formatting instructions.
- **System under test:** The same agent memory system processes all document sets (Section 4.1).
- **Scoring:** Same intrinsic scorer (claim coverage via embedding similarity + LLM entailment judge) and extrinsic scorer (QA accuracy).

For S1 and S2 we use a single high-capability model; for S3 (and the strategies built on it) we rotate across six models from five providers, one of which is that same single model; S1-lite uses a smaller model from the same family as the single model. The specific models are listed in Appendix B. All generation models differ from the system's extraction model, avoiding the stylistic bias that inflates scores when the same model generates and extracts [22].

3.4 Verification Pipeline

The verification pipeline (3-judge unanimous) is held constant for S1–S4 and the micro-ablations; S0 and S0-1judge explicitly vary verification as the experimental variable. For strategies S1–S6 and the micro-ablations, every generated document passes through 3

independent judge models from three providers (listed in Appendix B), following the multi-judge grounding methodology of [17, 19, 21]. The judge pool partially overlaps the generator pool, but all judges differ from the system's extraction model, so verification and scoring are not performed by the model under test, limiting self-preference effects [15]. All judges must unanimously agree that: (1) all target claims from `docClaims` are inferable from the document, (2) no extra factual information beyond the trajectory is introduced, and (3) the document reads naturally. Documents failing verification are regenerated up to 3 times with model rotation on failure.

S0 and S0-1judge relax this requirement, letting us measure both the direct effect of verification on document quality and its downstream impact on evaluation scores.

4 Experimental Setup

4.1 System Under Test

We evaluate a representative agent memory system following the standard architecture identified across recent systems [3, 9, 27]: (1) artifact ingestion, (2) LLM-based observation extraction producing atomic claims with provenance, (3) fact synthesis via embedding and top-K retrieval against existing facts, (4) LLM classification of the relationship (create, reinforce, update, supersede, none), and (5) fact store mutation. This architecture is representative of current agent memory systems, which despite surface-level differences in implementation all follow the same extract-retrieve-classify-store pattern [6]. The system is identical to the one evaluated in FEB-30 [19]. The system uses a single extraction model distinct from all generation models.

4.2 Metrics

We report both intrinsic and extrinsic metrics following FEB-30's evaluation framework [19]; the complementarity of these two evaluation modes is examined in [20]:

Intrinsic metrics:

- *Fact accuracy* (Acc): mean claim coverage across all trajectory states.
- *Update detection recall* (UDR): proportion of states requiring updates where the system actually modified its facts.
- *False update rate* (FUR): proportion of reinforce states where the system incorrectly modified its facts.
- *Classification accuracy* (CA): proportion of operations correctly classified.

Extrinsic metrics:

- *QA accuracy* (Extr): LLM-judged answer quality on trajectory-specific diagnostic questions.

Document characterization metrics (measured on generated documents, independent of the memory system):

- *Verbatim rate*: proportion of claims appearing word-for-word in the document.
- *Identifier preservation*: proportion of key identifiers (numbers, percentages, proper nouns) retained in the document.
- *Claim density*: target claims per 100 words (signal concentration).
- *Claim position*: mean normalized position of claims in the document (0=start, 1=end).
- *Vocabulary diversity*: per-document and corpus-level type-token ratio, corpus unique-vocabulary size, and length-robust lexical diversity (MATTR, moving-average type-token ratio with a 100-token window; MTLD, measure of textual lexical diversity; and distinct-bigram ratio).
- *Structural metrics*: headings, bullet points, tables, code blocks, and signatures per document.
- *Readability*: Flesch-Kincaid grade level and reading ease score.
- *Verification metrics*: first-pass verification rate, retry distribution per model, split verdict rate (proportion of documents receiving non-unanimous judge decisions).

Each strategy is scored 3 times to account for LLM scorer variance. We report means and standard deviations across these runs; the across-run standard deviation is the score-variance metric. For significance, we pre-specify 7 ladder-adjacent contrasts (each isolating one component: S0–S1, S1–S2, S2–S3, S3–S4, plus S1 against the two micro-ablations S1-lite and S1-nonoise, and S1–S3) and assess them via paired t-tests over the 30 per-trajectory fact-accuracy values with Bonferroni correction over these 7 contrasts. We restrict to these adjacent contrasts because the ladder is designed so that each isolates a single component.

5 Results and Analysis

5.1 Overall Scores by Strategy

Table 2 shows overall evaluation scores across generation strategies. S3 (multi-model rotation) attains the highest fact accuracy (64.8%) with low across-run variance (±0.5%), among the most stable of the strategies. S1-lite (smaller generator) produces the lowest scores across most metrics, consistent with generator capability affecting benchmark difficulty. The verification ablation shows that S0 (no verification) attains comparable fact accuracy to S1 but with lower update detection recall (78.7% vs. 86.6%), indicating that unverified documents more frequently omit claims. Across the 7 pre-specified contrasts (paired t-tests, Bonferroni-corrected), the realism contrast (S3 vs. S4) is already significant at n=30 ($p < 0.05$; driven by L3). The remaining contrasts do not reach significance at this sample size but may do so on a larger benchmark. Notably, extrinsic QA accuracy stays nearly flat across strategies (46.7–53.3%) while intrinsic fact accuracy spans 8.1 pp, so downstream QA is insensitive to the generation-strategy differences that intrinsic inspection reveals.

5.2 Difficulty Calibration

Table 3 shows the difficulty-level breakdown per strategy. At face value most single-model strategies show an L1/L2 inversion, but these per-level differences are within generation variance: strategies that share an L1/L2 generation procedure (e.g. S3, S4, S5) nonetheless differ by 7.8 pp at L1 and 11.1 pp at L2 across their independent generations, comparable to or larger than the differences between strategies at a given level. Many apparent inversions fall within this noise (e.g. S3's L1 = 71.9 vs. L2 = 72.1). The one robust pattern is the L3 drop, which is large (15–25 pp below L1/L2) and consistent across all strategies; L3 also carries the lowest across-run std of the three levels, indicating that L3 difficulty (requiring inference) is the

Strategy	Fact Acc (%)	UDR (%)	FUR (%)	Class. Acc (%)	QA Acc (%)
Verification ablation					
S0 (no verification)	63.6 ± 0.4	78.7 ± 3.2	18.3 ± 1.6	18.6 ± 2.2	51.7 ± 0.0
S0-1judge	62.0 ± 2.0	83.8 ± 1.3	18.7 ± 3.0	15.2 ± 0.3	53.3 ± 1.4
Main ladder					
S1 (single model)	63.1 ± 2.7	86.6 ± 2.2	18.8 ± 1.1	13.7 ± 1.0	50.0 ± 1.4
S2 (+ style)	61.9 ± 0.5	85.2 ± 2.1	18.2 ± 1.5	18.6 ± 1.4	50.0 ± 3.6
S3 (+ multi-model)	64.8 ± 0.5	89.7 ± 1.5	19.2 ± 1.4	17.9 ± 1.9	50.0 ± 2.4
S4 (+ hybrid)	56.7 ± 0.8	93.1 ± 0.5	21.5 ± 1.5	18.6 ± 2.4	49.4 ± 2.1
S5 (+ human review)	59.0 ± 1.6	93.8 ± 1.7	21.0 ± 1.7	17.2 ± 2.4	48.3 ± 2.4
S6 (+ sequential)	60.6 ± 1.6	93.5 ± 1.9	21.5 ± 1.3	20.0 ± 1.3	50.6 ± 2.8
Micro-ablations					
S1-lite (smaller model)	59.7 ± 1.4	83.5 ± 1.5	20.3 ± 0.9	18.6 ± 0.8	46.7 ± 1.4
S1-nonoise (no distractors)	60.1 ± 2.8	86.3 ± 1.3	21.9 ± 1.1	14.1 ± 1.3	48.3 ± 2.4
S4-persona	61.0 ± 0.5	94.2 ± 2.6	21.1 ± 2.4	16.5 ± 0.4	47.2 ± 0.8

Table 2: Overall evaluation scores by generation strategy (mean ± std across 3 scoring runs). UDR = update detection recall, FUR = false update rate (lower is better); for the other metrics higher is better.

Strategy	L1	L2	L3
S0	70.4 ±0.6	67.2 ±0.6	53.1 ±0.2
S0-1judge	64.1 ±6.0	68.0 ±5.5	53.8 ±2.2
S1	66.6 ±1.9	68.6 ±6.4	53.9 ±2.2
S1-lite	57.3 ±4.3	67.1 ±1.2	54.8 ±2.9
S1-nonoise	62.5 ±8.2	66.9 ±3.9	50.8 ±3.0
S2	65.9 ±1.7	64.4 ±2.0	55.5 ±1.4
S3	71.9 ±2.7	72.1 ±3.1	50.3 ±2.2
S4	64.1 ±3.6	61.0 ±5.3	44.9 ±2.1
S4-persona	62.8 ±3.9	72.2 ±1.8	48.0 ±1.1
S5	67.4 ±3.9	64.1 ±3.3	45.6 ±2.7
S6	63.5 ±2.8	70.5 ±4.5	47.9 ±2.2

Table 3: Fact accuracy (%) by difficulty level per strategy (mean ± std across 3 runs). L3 has the lowest across-run std at nearly every strategy; L1/L2 differences are within generation variance.

one difficulty signal that survives generation variance. Generation strategy also shifts per-category accuracy (up to 21.5 pp on the classification category), confirming that strategy affects not just overall scores but where a benchmark looks hard. We report the full breakdown in Appendix C.

5.3 Generation Pipeline Characterization

Several patterns emerge from generation-time and document-level metrics in Table 4. Verbatim reproduction of claims is near-zero across all strategies (≤0.8%), confirming that the generation prompt successfully enforces paraphrasing. Readability (Flesch-Kincaid grade level) is largely consistent across strategies (12.8–13.7, college-level text), with the exception of S1-nonoise which produces denser technical prose (grade 14.1, ease score 23.9 vs. 28–30 for other strategies) without conversational filler to break up the content. Notably, removing noise does not reduce document length: S1-nonoise averages 436 words vs. S1's 418 words, as the model compensates for missing noise by expanding on-topic discussion to meet the prompted length target.

Structural analysis reveals that single-model strategies (S0/S1) produce highly formulaic documents averaging 9–10 headings per document with signatures in 84–98% of documents but no tables. Multi-model rotation (S3) introduces structural diversity: fewer headings (6.2), tables in 35% of documents, and signatures in only 72%. S4/S6 (hybrid injection, sequential) have the least structured output (3.6–4.1 headings, 64–65% signatures) because they incorporate real human text which is inherently less templated. This structural homogeneity in single-model strategies means a memory system tuned for heading-delimited extraction would score artificially high on S1 but fail on S4, where claims are embedded in unstructured prose.

Multi-model strategies (S3, S4-persona, S6) produce longer documents (489–592 words) with lower claim density (0.33–0.40 claims/100 words) compared to single-model strategies (414–436 words, 0.45–0.47 claims/100 words). This means claims are more diluted in multi-model documents, requiring better extraction capabilities. Vocabulary diversity reinforces this gap: multi-model corpora contain nearly double the unique vocabulary (5,700–6,300 types) compared to single-model corpora (3,400–3,750 types), with per-document type-token ratios of 0.57–0.58 vs. 0.52–0.53. Notably, style grounding alone does not increase vocabulary (S1: 3,451 types → S2: 3,503 types), while model rotation nearly doubles it (S2 → S3: 6,205 types), and adding explicit author personas pushes it further (S4-persona: 6,295 types), confirming that lexical diversity is driven by generator diversity and persona differentiation, not by the style reference corpus. A memory system's embedding model must therefore handle substantially more lexical variation under multi-model generation. Table 5 summarizes these metrics.

Identifier preservation is highest for the multi-model strategies (S3 64.0%, S4-persona 60.8%, S6 59.4%) and lowest for the unverified single-model strategies (S0 49.1%, S0-1judge 54.8%); the smaller single model S1-lite is an exception, matching S4-persona at 60.8%. This indicates that verification and model diversity help retain exact numbers and names rather than paraphrasing around them, which

Strategy	W/d	Att.	P%	Hdg	Bul	Tbl%	Cd%	Sig%	Verb%	ID%	Cl/100w	Pos
S0	414	1.00	100	9.4	9.0	0	1	84	0.0	49.1	0.47	0.25
S0-1judge	417	1.09	98	9.4	8.8	0	1	88	0.0	54.8	0.47	0.25
S1	418	1.20	98	9.7	8.8	0	1	86	0.0	51.2	0.47	0.26
S1-lite	496	1.09	98	9.9	10.8	0	5	98	0.4	60.8	0.39	0.25
S1-nonoise	436	1.13	98	9.5	10.8	1	0	82	0.0	56.9	0.45	0.27
S2	419	1.23	96	8.1	8.9	0	1	89	0.0	58.0	0.47	0.26
S3	592	1.12	98	6.2	10.8	35	3	72	0.0	64.0	0.33	0.28
S4	506	1.03	100	4.1	6.8	22	0	64	0.4	57.2	0.39	0.35
S4-persona	526	1.06	99	3.6	7.3	19	9	65	0.0	60.8	0.37	0.31
S6	489	1.07	100	3.9	6.7	29	3	65	0.8	59.4	0.40	0.36

Table 4: Generation pipeline and document-level metrics. W/d = mean words per doc. Att. = mean generation attempts before verification passes. P% = verification pass rate (fraction eventually verified within 3 attempts). Hdg/Bul = mean headings/bullets per doc. Tbl/Cd/Sig = % of docs containing tables, code blocks, or signatures. Verb% = claims appearing verbatim. ID% = identifier preservation rate (key numbers/names retained). Cl/100w = target claims per 100 words (signal density). Pos = mean claim position in document (0=start, 1=end).

Strategy	MATTR	MTLD	Dist-2	Doc TTR	Vocab	Corpus TTR
S0	0.731	108.1	0.882	0.532	3,466	0.059
S0-1judge	0.730	106.3	0.882	0.531	3,343	0.057
S1	0.732	108.0	0.885	0.529	3,451	0.058
S1-lite	0.734	107.2	0.880	0.516	3,722	0.053
S1-nonoise	0.722	100.5	0.873	0.518	3,754	0.062
S2	0.729	106.8	0.885	0.533	3,503	0.059
S3	0.788	160.1	0.907	0.568	6,205	0.079
S4	0.766	141.3	0.900	0.575	6,015	0.084
S4-persona	0.773	148.9	0.903	0.572	6,295	0.086
S6	0.771	135.5	0.903	0.583	5,688	0.083

Table 5: Vocabulary diversity metrics. MATTR = moving-average type-token ratio (window 100). MTLD = measure of textual lexical diversity (both length-robust). Dist-2 = distinct bigram ratio. Doc TTR = mean per-document type-token ratio. Vocab = unique word types across the corpus. Corpus TTR = types/tokens across all documents. MATTR, MTLD, and Dist-2 use lowercased alphanumeric tokens; higher is more diverse for all columns.

makes extraction easier for identifier-based claims. S4 and S6 push claims later in documents (position 0.35–0.36 vs. 0.25 for S1), reflecting that hybrid injection and sequential context embed claims in the middle or end rather than frontloading them. S2 (style-grounded) has the lowest first-pass verification rate (96.1%), indicating that style references sometimes lead the model to prioritize mimicking the reference over encoding all target claims.

Because raw vocabulary counts scale with corpus length, we also report MATTR, which measures lexical diversity within a fixed 100-token window and is therefore length-independent. The multi-model strategies (S3, S4, S4-persona, S6: MATTR 0.766–0.788) remain more diverse than the single-model strategies (S0–S2, S1-lite, S1-nonoise: 0.722–0.734), confirming that the diversity gap is not an artifact of the longer documents multi-model rotation produces. MTLD, a second length-robust measure with a different mechanism, shows a wider separation in the same direction (135–160 vs. 100–108), and distinct-bigram ratio agrees (0.900–0.907 vs. 0.873–0.885).

5.4 Model-Specific Generation Biases

In S3, where 6 models rotate, per-model verification statistics reveal systematic differences. Average retry counts ranged from 1.0 to 1.23 attempts across the pool, with most models passing on the first attempt. These differences are modest with identifier-preservation prompting in place, but prior to this mitigation, one candidate model exhibited 40–50% verification failure rates due to identifier abstraction (replacing "99.95%" with "very high level of availability, consistent with industry standards" and "Q3 2025" with "the third quarter"). Notably, that model was removed from the generation rotation after exhibiting these high abstraction rates, yet abstraction tendency did not predict a model's ability to verify claim presence in others' documents, illustrating that generation and verification are independent capabilities.

Manual review of the 127 generated documents across S4 revealed 9 documents (7.1%) with misalignment between content and trajectory ground truth, classified into four subtypes: (1) *forward leakage* (3 cases): early-state documents reference information from later states because the model infers the trajectory's conclusion from the topic description; (2) *premature decisiveness* (2 cases): tentative claims rendered as firm decisions; (3) *contradicting details*

(3 cases): model fabricates specifics not in the claims; (4) *terminology mismatch* (1 case): model substitutes service names with paraphrases. These failure modes are complementary across models, further motivating rotation as a mitigation strategy.

The 3-judge unanimous verification is effective at catching missing claims but cannot detect forward leakage (judges do not have access to future states), cross-document contradictions (judges evaluate documents independently), numeric/terminology drift where judges accept semantic equivalents (e.g., "dozens" ≈ "47"), or style-reference bleed-through, where off-topic content from the style exemplar (meeting-room design, prototype scoring) is carried into the generated document and accepted as plausible in-domain material. Because the judges check claim presence rather than surface form, they likewise do not flag non-semantic artifacts, template placeholders, residual scaffolding, or non-inclusive terminology, which require a separate lint pass to remove. This explains why S0 (no verification) achieves similar fact accuracy to S1: verification primarily catches claim *absence*, not claim *distortion*. The 8 pp UDR difference between S0 and S1 reflects cases where unverified documents genuinely omit claims. Comparing S0-1judge to S1 reveals that single-judge verification is surprisingly close on aggregate (62.0% vs. 63.1%), but with higher variance (±2.0 vs. S3's ±0.5). Approximately 12% of documents received a non-unanimous (split) verdict on first generation; these are regenerated (up to 3 attempts) rather than accepted, so the first-pass rate is around 87–93% for the verified single-generation strategies even though the eventual pass rate (P% in Table 4) reaches 96–100%. Unanimous consensus thus filters borderline cases, producing a more consistently calibrated document set.

Approximately 28% of generated documents (36/127) required post-processing to remove scaffolding artifacts (code fences, preamble commentary). This rate was model- and document-type-specific: one model produced code fences on 100% of email-type documents, another added preamble on 100% of hybrid injection documents, a third on ∼14% of status updates, while the remaining three produced clean output. All artifacts are removed automatically. In S3, rotation distributes these artifacts; in S1/S2, they are uniformly present or absent. Finally, S1-lite generates longer documents (496 words vs. S1's 418 words) yet scores lower on fact accuracy (59.7% vs. 63.1%), as the additional length comes from filler that dilutes signal density. Conversely, S3's longer documents (592 words) come from stylistically richer content that provides better extraction context.

5.5 Style Grounding Effects

Comparing S1 (63.1%) vs. S2 (61.9%) shows that style grounding produces a small decrease in fact accuracy (−1.2 pp) while reducing across-run variance (from ±2.7 to ±0.5). Style references make documents more realistic but slightly harder to extract from, while making scores more stable across runs. The accompanying change in classification accuracy (13.7% to 18.6%) is not statistically significant (paired t, $p = 0.13$).

5.6 Multi-Model Rotation as Bias Averaging

S3 achieves the highest fact accuracy (64.8%) and low across-run variance (±0.5%) despite individual models in the pool having different strengths. This is consistent with an ensemble effect: each model introduces its own bias profile (e.g., forward leakage, residual scaffolding, or verbosity without precision), and rotation averages across them, producing an evaluation signal that is stable across scoring runs.

The update detection recall improvement (S1: 86.6% → S3: 89.7%) suggests that multi-model stylistic diversity produces documents where factual updates are expressed differently enough from prior states that the system detects them as changes rather than reinforcements. S6 (sequential context) achieves the highest C5 (temporal correctness) score across all strategies (65.9%), indicating that when the generator sees prior documents, it produces temporal evolution that is more clearly expressed and easier for the memory system to track correctly.

Identifier preservation follows the same logic: a single model that abstracts specifics degrades the whole corpus, whereas under rotation its effect is diluted by the models that abstract less, so the multi-model strategies retain exact numbers and names best (Table 4). Combined with 3-judge verification, this gives S3 the strongest fidelity on exact-identifier claims among the automated strategies. Human review (S5) can add further correctness by catching errors automated verification misses, but at substantial human-time cost.

5.7 Cost-Benefit Analysis

Table 6 shows token-weighted costs computed from actual prompt content and model-specific pricing. Judge verification dominates, accounting for 60–67% of total spend. Coupling cost to a robust document property is more informative than raw spend: multi-model rotation (S3) roughly doubles corpus vocabulary diversity (3,451 → 6,205 unique types) over single-model generation (S1) for only ∼7% additional cost, and the verified realism-oriented strategies (S3, S4, S4-persona) attain the highest vocabulary-to-cost score (V/C) among verified strategies. The length-normalized variant (Q/C) gives the same ordering among verified strategies, so the diversity-per-cost advantage is not an artifact of these strategies producing longer documents. The σ·\$ column combines across-run stability and cost (lower is better on both); S4 is in fact slightly cheaper than S1 (0.90×) because near-perfect first-pass verification means fewer retries.

5.8 Robustness to Regeneration

We first vary the system under test while holding the generated corpus fixed. Re-scoring the original corpus on the same system after minor upgrades (three scoring runs per cell) shifts fact accuracy only slightly (S0 $63.6 \pm 0.4 \rightarrow 62.1 \pm 2.3$, S1 $63.1 \pm 2.7 \rightarrow 63.9 \pm 1.3$, S3 $64.8 \pm 0.5 \rightarrow 66.4 \pm 1.8$; at most 1.6 pp in magnitude). System upgrades of this scale therefore move measured accuracy less than the generation-strategy differences we study.

We then hold the system fixed and vary the generated corpus, scoring three independently generated corpora per strategy (the original corpus plus two regenerations, three scoring runs each). The between-corpus standard deviation of fact accuracy is 2.5 pp for S1, 1.5 pp for S3, and 1.0 pp for S4, so the single-model baseline is the most regeneration-sensitive and the two realism strategies the most stable. These between-corpus values exceed the scorer-run std in Table 2 (0.5–1 pp) but remain well below the largest strategy

	Generation (per doc)				Verification (per doc)							
Strategy	**Calls**	**In**	**Out**	**$gen**	**Calls**	**In**	**Out**	**$jdg**	**Rel.$**	**σ·$**	**V/C**	**Q/C**
S0	1.00	262	537	.245	0	0	0	.000	0.31×	0.12	1.04	1.14
S0-1judge	1.09	286	596	.271	1.09	752	88	.149	0.53×	1.06	0.71	0.83
S1	1.20	314	655	.298	3.61	2481	289	.491	1.00×	2.70	0.13	0.24
S1-lite	1.09	287	704	.024	3.28	2582	263	.495	0.66×	0.92	0.67	0.72
S1-nonoise	1.13	265	640	.287	3.40	2406	272	.473	0.96×	2.69	0.28	0.14
S2	1.23	503	680	.327	3.69	2563	295	.506	1.06×	0.53	0.07	0.12
S3	1.12	468	861	.278	3.35	3064	268	.569	1.07×	0.54	0.97	1.00
S4	1.03	620	675	.240	3.09	2470	248	.472	0.90×	0.72	1.13	0.89
S4-persona	1.06	664	711	.254	3.17	2589	253	.492	0.95×	0.47	1.16	0.93
S6	1.07	1137	686	.289	3.21	2519	257	.483	0.98×	1.57	0.91	0.86

Table 6: Token-based cost analysis (127 docs total). In/Out = mean input/output tokens per document, computed from actual prompt content and generated documents. $gen/$jdg = total USD cost per corpus using model-specific pricing. Rel.$ = total cost relative to S1. σ·$ = across-run accuracy std × relative cost (lower is better on both axes). V/C = normalized vocabulary-to-cost score, the sum of min–max normalized vocabulary diversity and inverse cost (range 0–2, higher is better); note that the unverified S0 scores high only through its low cost, not document quality. Q/C uses the same construction but with length-normalized MATTR in place of raw vocabulary, so it rewards per-length lexical diversity rather than document length. Output tokens include retried generations (paid but discarded). S5 is omitted as its added cost is human-review time rather than model tokens. See Appendix A for full calculation methodology.

effects (up to 8.1 pp), and the ordering we rely on is preserved: across all three corpora the multi-model and realism strategies stay above the single-model baseline (between-corpus means S3 64.8 ± 1.5, S4 63.2 ± 1.0 vs. S1 61.1 ± 2.5). For a fixed benchmark, then, the generated documents influence measured accuracy more than minor system upgrades do, and the generation strategy more still, so generation is a first-order experimental factor rather than a neutral substrate.

Regeneration also clarifies the S3–S4 comparison. S3 and S4 share their L1/L2 generation and differ only in the L3 hybrid injection, and across the regenerated corpora they are close at L1/L2 (S3 ≈ 66/74, S4 ≈ 71/69) so their overall scores converge (Section 5.8 between-corpus means S3 64.8, S4 63.2). The difference between the two strategies is concentrated at L3, where hybrid injection lowers S4's accuracy below S3's on both the original corpus (Table 3, S3 50.3 vs. S4 44.9) and the regenerated corpora (S3 ≈ 54 vs. S4 ≈ 49), a 5–7 pp gap in the same direction. This lower L3 score is not evidence that S4 is a weaker strategy: injecting claims into real host documents makes L3 harder and closer to real-world extraction, so a lower score here reflects a more demanding benchmark rather than a worse one. The larger overall gap on the original corpus (56.7 vs. 64.8) reflected additional L1/L2 regeneration variance on top of this L3 effect. L3 is also the most stable level across runs (regenerated-corpus std ≤ 0.5 pp vs. up to 6 pp at L1/L2), so the L3 contrast is the most reliable one.

While scores move under regeneration, the document-characterization metrics do not. Computing the Section 5.3 metrics on three independently generated corpora (the original plus two regenerations) for all ten automated strategies leaves the characterization patterns intact: length-normalized lexical diversity is essentially unchanged (across regenerations the MATTR standard deviation is ≤ 0.01 and the distinct-bigram ratio ≤ 0.01 for every strategy, and MTLD preserves the same separation, keeping the multi-model vs. single-model diversity gap intact in each corpus), generation effort is stable (mean attempts std ≤ 0.05, verification pass-rate std ≤ 1 pp), claim density and claim position are invariant, and the identifier-preservation ordering holds (multi-model strategies highest, unverified single-model lowest, with a between-corpus standard deviation of 0.3–3.7 pp). The document-level properties the characterization analysis relies on are therefore attributes of the generation strategy rather than of a single generated corpus, even though the system scores computed from those documents are more regeneration-sensitive.

6 Discussion

A generation strategy is not judged by how high the system scores on it; a higher score can indicate a worse benchmark. We treat a strategy as sound when its documents are (i) *correct*, i.e. they actually encode the intended claims, and (ii) *appropriately difficult*, i.e. realistic rather than formulaic. Both can fail independently of the score. Unverified strategies (S0, and to a lesser extent single-judge S0-1judge) risk documents that do not reliably encode the target claims at all, since no judge confirms their presence; conversely, simpler or homogeneous strategies (S0, S1-lite) can be too easy, producing structurally simple, formulaic documents that inflate scores without reflecting real-world difficulty. A system scoring 63.6% on S0-generated documents may therefore score substantially lower on authentic multi-author corpora. S4 (hybrid injection) and S5 (human-reviewed) produce the lowest fact accuracy scores (56.7%, 59.0%); we attribute this to their documents being closest to real-world conditions, but lower accuracy is not by itself evidence of greater realism and must be read alongside the document-characterization metrics (Section 5.3), since a strategy could instead lower scores by introducing uncontrolled difficulty. Practitioners should interpret benchmark scores relative to the generation strategy used: high scores on simple strategies indicate the system works under idealized conditions, not that it is production-ready.

Generation strategy matters most when: (1) the evaluation target is fine-grained (specific identifier extraction vs. topic-level recall), (2) the benchmark includes multiple difficulty levels that require calibration, and (3) the comparison involves systems with different extraction architectures. For coarse evaluations ("does the system remember anything about this topic?"), even S1 may suffice.

Based on our results, we recommend:

(1) Use verified multi-model rotation (S3) as the default for published benchmarks: verification guards document correctness, and model diversity avoids the formulaic homogeneity that makes single-model strategies too easy.
(2) Use hybrid injection (S4) or human review (S5) when realism matters, as these most closely approximate real-document difficulty.
(3) Use style-grounded single-model (S2) for rapid internal iteration when rankings are not critical.
(4) Always verify identifier preservation. A generator that abstracts specifics produces documents that no longer faithfully encode the target claim, regardless of whether this raises or lowers measured scores.
(5) Report the generation strategy alongside benchmark results to enable reproducibility assessment.

This study uses a single system under test, so we cannot directly show that generation strategy changes the relative ranking of different memory systems; our category-ranking analysis is a proxy for this effect, not a demonstration of it. We do find, however, that scores are stable to a within-system upgrade (Section 5.8), so the reported differences are not an artifact of one specific system build. The system's classification step moreover operates near chance (13.7–20.0% for a 5-way decision). The per-strategy ±std in Table 2 reflects scorer-run variability on a single corpus. To test stability across regeneration we also scored multiple independently generated corpora per strategy (Section 5.8) and find the between-corpus variance larger than scorer-run variance but below the main strategy effects, with the strategy ordering preserved. The strategy axes are also not fully orthogonal: multi-model rotation (S3) changes both model diversity and the specific generator capabilities, hybrid injection (S4) changes both the content source and the injection mechanism, and human review (S5) changes the validation process, so observed differences cannot always be attributed to a single named factor. Additionally, the 3-judge verification ensures claim presence but does not fully control for difficulty variation across strategies, as documents can pass verification while being easier or harder to extract from. Both the verification judges and the scoring judges are themselves LLMs, so the measured scores may partly reflect judge behavior rather than generation strategy alone. The specific score and cost numbers are tied to our particular model pool and trajectory design, and since the benchmark, scoring, and system under test are those of FEB-30, some recommendations echo its design choices. A further limitation is that a curated generated corpus under-samples the sparse, partial inputs a live pipeline ingests: generated documents tend to carry fuller, more uniform entity detail than real-world artifacts, so generation strategy also determines which slice of the input distribution the benchmark exercises. Finally, editing the corpus for public release left aggregate scores essentially unchanged, consistent with the corpus-version robustness reported for FEB-30 [19].

This work uses trajectories with fictional entities and synthetically generated documents, plus publicly available research corpora [2, 8]. The Enron corpus contains real individuals' names and communications; we use it only under its public release terms as a stylistic/host source, and the publicly released FEB-30 corpus has source-corpus personal information redacted (no new personal data is collected). The broader impact is positive: improving evaluation methodology reduces the risk of deploying agent memory systems based on misleading benchmark results.

7 Conclusion

We present, to our knowledge, the first empirical study of how synthetic document generation strategy affects intrinsic agent memory evaluation, enabled by the intrinsic-evaluation substrate of FEB-30. Ablating over 11 strategies spanning 5 axes (verification strictness, generator size, style grounding, model diversity, and document realism) on the same 30 fact-evolution trajectories with identical ground truth, we show that generation strategy introduces measurable differences in evaluation outcomes. Fact accuracy varies by up to 8.1 percentage points across strategies, and document-level characteristics (vocabulary diversity, claim density, structural form) differ markedly by strategy. Multi-model rotation (S3) achieves the highest accuracy and low across-run variance at only 7% additional cost, while hybrid injection (S4) and human review (S5) best approximate real-document difficulty.

As agent memory systems move from research to deployment, evaluation methodology must be trustworthy. Our structural findings (multi-model rotation reduces variance and averages out model-specific identifier abstraction, style grounding stabilizes output, verification catches absence but not distortion) may generalize to other benchmarks and model pools, as they appear to reflect architectural properties rather than model-specific behavior. The specific score differences and cost ratios are contingent on our model pool and trajectory design. We provide concrete guidelines for practitioners and recommend that generation strategy be treated as an explicit methodological choice to be reported and justified.

References

[1] Amazon Artificial General Intelligence. 2025. The Amazon Nova Family of Models: Technical Report and Model Card. arXiv:2506.12103. doi:10.48550/arXiv.2506.12103
[2] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, and Melissa Kronenthal. 2005. The AMI Meeting Corpus: A Pre-Announcement. In *Proc. MLMI Workshop*.
[3] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *arXiv preprint arXiv:2504.19413* (2025). doi:10.48550/arXiv.2504.19413
[4] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437. doi:10.48550/arXiv.2412.19437
[5] Yuanzhe Hu, Yu Wang, and Julian McAuley. 2026. Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions. In *ICLR*. doi:10.48550/arXiv.2507.05257
[6] Dongming Jiang, Yi Li, Songtao Wei, Jinxin Yang, Ayushi Kishore, and Alysa Zhao. 2026. Anatomy of Agentic Memory: Taxonomy and Empirical Analysis of Evaluation and System Limitations. *arXiv preprint arXiv:2602.19320* (2026). doi:10.48550/arXiv.2602.19320
[7] Seokwon Jung, Alexander Rubinstein, Arnas Uselis, Sangdoo Yun, and Seong Joon Oh. 2026. MEME: Multi-Entity and Evolving Memory Evaluation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.*

doi:10.48550/arXiv.2605.12477
[8] Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Proc. European Conference on Machine Learning (ECML)*.
[9] Christopher Latimer, Nicolò Boschi, Andrew Neeser, Ethan Peng, and Saransh Gaddipati. 2026. Hindsight: Structured Agent Memory that Retains, Recalls, and Reflects. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics. doi:10.18653/v1/2026.acl-demo.27
[10] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. 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)*. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.747
[11] Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of Synthetic Data as a Benchmark. *arXiv preprint arXiv:2409.11968* (2024). doi:10.48550/arXiv.2409.11968
[12] Mistral AI. 2024. Large Enough (Mistral Large 2). https://mistral.ai/news/mistral-large-2407/.
[13] Mistral AI. 2025. Magistral. arXiv:2506.10910. doi:10.48550/arXiv.2506.10910
[14] Mistral AI. 2025. Mistral 3. https://mistral.ai/news/mistral-3/.
[15] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *NeurIPS*. doi:10.48550/arXiv.2404.13076
[16] Hossein A. Rahmani, Varsha Ramineni, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2025. Towards Understanding Bias in Synthetic Data for Evaluation. In *Proceedings of CIKM*. doi:10.48550/arXiv.2506.10301
[17] Joan Vinyals Roig. 2025. Scalable and Reliable Evaluation of AI Knowledge Retrieval Systems: RIKER and the Coherent Simulated Universe. *arXiv preprint arXiv:2601.08847* (2025). doi:10.48550/arXiv.2601.08847
[18] Max Schaffelder and Albert Gatt. 2026. Synthetic Eggs in Many Baskets: The Impact of Synthetic Data Diversity on LLM Fine-Tuning. In *Findings of the Association for Computational Linguistics: ACL 2026*. doi:10.18653/v1/2026.findings-acl.360
[19] Sebastian-Antonio Toma, Jean de Kernier, Mike Ply, and Oliviu Matei. 2026. Beyond Downstream QA: Rethinking How We Evaluate Agent Memory. arXiv:2610.22209 [cs.CL] https://arxiv.org/abs/2610.22209
[20] Sebastian-Antonio Toma and Oliviu Matei. 2026. What Agent Memory Benchmarks Miss: The Gap Between Response Quality and Knowledge State Correctness. In *Proceedings of InnoComp 2026 (Communications in Computer and Information Science)*. Springer. To appear.
[21] Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. 2026. From Recall to Forgetting: Benchmarking Long-Term Memory for Personalized Agents. In *Findings of the Association for Computational Linguistics: ACL 2026*. doi:10.18653/v1/2026.findings-acl.1337
[22] Jonas van Elburg, Peter van der Putten, and Maarten Marx. 2026. Can We Evaluate RAGs with Synthetic Data?. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Irena Koprinska, João Mendes-Moreira, and Paula Branco (Eds.). Springer Nature Switzerland, Cham, 177–192. doi:10.1007/978-3-032-19102-1_11
[23] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *International Conference on Learning Representations (ICLR)*. doi:10.48550/arXiv.2410.10813
[24] Lingze Zhang and Ellie Pavlick. 2025. Does Training on Synthetic Data Make Models Less Robust?. In *Proceedings of the Sixth Workshop on Insights from Negative Results in NLP*. doi:10.18653/v1/2025.insights-1.8
[25] Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, and Jishen Zhao. 2026. AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications. In *ICML*. doi:10.48550/arXiv.2602.22769
[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks*. doi:10.48550/arXiv.2306.05685
[27] Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Jie Li, Wen-Ji Dong, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. 2026. AriadneMem: Threading the Maze of Lifelong Memory for LLM Agents. *arXiv preprint arXiv:2603.03290* (2026). doi:10.48550/arXiv.2603.03290

A Cost Calculation Methodology

A.1 Token Estimation

Input and output tokens are estimated using a word-to-token ratio of 1.3 (standard for English text with BPE tokenizers). For each document:

- **Generation input tokens** are reconstructed from actual prompt content: the base prompt template (measured at 204 tokens), actual claim text from the trajectory definitions, noise claims from the shared pool (for L2/L3), style reference text from the curated AMI [2] and Enron [8] documents (measured average 269 tokens after truncation at 1500 characters), hybrid host document text (measured average 441 tokens), persona prefix (measured 40 tokens), and sequential context from prior generated documents (each truncated at 1500 characters).
- **Generation output tokens** = word count of the saved document × 1.3, multiplied by the number of generation attempts (retried generations produce output tokens that are paid but discarded).
- **Judge input tokens** = judge prompt template (113 tokens) + generated document tokens + claim text tokens, multiplied by number of judges × attempts.
- **Judge output tokens** = 80 tokens per judge call (the VERDICT/MISSING/EXTRA/FEEDBACK response format), multiplied by judges × attempts.

A.2 Model Pricing

USD costs are computed using published per-token pricing (per 1M tokens):

- **S0, S0-1judge, S1, S1-nonoise, S2** (single high-capability generator): input \$0.80, output \$3.20.
- **S1-lite** (smaller, cheaper generator from the same family): input \$0.06, output \$0.24 (13× cheaper).
- **S3, S4, S4-persona, S6** (weighted average of 6 rotation models): input \$0.69, output \$2.17. Computed as the mean of the six models' input/output prices (per 1M tokens): \$0.80/\$2.40, \$2.00/\$6.00, \$0.15/\$0.20, \$0.80/\$3.20, \$0.30/\$0.90, \$0.10/\$0.30.
- **Verification judges** (average of 3 judge models): input \$1.17, output \$3.31. Computed as mean of three models with pricing ranging from \$0.72–\$2.00 per 1M input tokens and \$0.72–\$6.00 per 1M output tokens.

A.3 Derived Metrics

- **\$gen** = (total generation input tokens × gen input price + total generation output tokens × gen output price) / 1,000,000, for all 127 documents.
- **\$jdg** = (total judge input tokens × judge input price + total judge output tokens × judge output price) / 1,000,000, for all 127 documents.
- **Rel.\$** = (\$gen + \$jdg) for this strategy / (\$gen + \$jdg) for S1.
- σ·\$ = across-run accuracy standard deviation (from 3 scoring runs) × Rel.\$. Lower values indicate a strategy that is both stable across runs and inexpensive.
- **V/C** = normalized vocabulary-to-cost score. We min–max normalize corpus vocabulary size to $[0, 1]$ (higher is better)

Category	S0	S0-1j	S1	S1-lite	S1-nn	S2	S3	S4	S4-p	S5	S6	Δ
C1	69.1	70.1	80.9	61.6	71.9	66.4	78.4	77.1	66.7	80.9	68.4	19.3
C2	64.9	59.4	60.3	62.2	61.4	66.7	59.6	56.3	59.1	55.9	56.1	10.8
C3	82.6	78.5	75.0	74.3	81.9	81.9	87.5	66.0	81.3	67.4	74.3	21.5
C4	52.0	62.7	60.4	65.1	58.3	51.9	61.1	51.6	53.9	50.4	58.6	14.7
C5	61.1	54.4	57.5	55.4	50.2	56.0	61.3	50.8	63.2	58.7	65.9	15.7
C6	57.8	49.7	47.2	41.7	41.4	54.7	46.1	41.1	47.2	44.2	43.6	16.7

Table 7: Fact accuracy (%) by diagnostic category per strategy (mean across 3 runs). Δ = max − min across all strategies. S0-1j = S0-1judge, S1-nn = S1-nonoise, S4-p = S4-persona. Categories differ in trajectory count (C3 has 4, C4 has 6, the others 5), so averaging the category means does not reproduce the trajectory-weighted overall accuracy in Table 2.

and relative cost to $[0, 1]$ (inverted, so lower cost scores higher), then sum them (range 0–2). Normalization is over the strategies in Table 6.

- **Q/C** = quality-to-cost score. Identical construction to V/C, but the quality term is the length-normalized MATTR (Table 5) in place of raw corpus vocabulary size, so the score rewards per-length lexical diversity rather than document length. Normalization is over the same strategies.

A.4 Data Sources

- Generation attempts and verification outcomes: recorded alongside each document set.
- Document word counts: measured from the saved documents.
- Claim text: from the trajectory definitions (per-document claims when present, otherwise the cumulative state claims).
- Style references: from curated AMI abstractive summaries [2] and Enron emails [8] (12 each), truncated at 1500 characters.
- Model pricing: from published model provider pricing pages as of July 2026.

B Generation Models

The single-model strategies (S1, S2, and the verification and noise ablations) use Amazon Nova Pro [1]. S1-lite uses Amazon Nova Lite v1 [1], a smaller, cheaper model from the same family as Nova Pro. The multi-model rotation (S3, S4, S4-persona, S6) rotates across six models from five providers (the two unnamed models below are each from a distinct provider not shared with any other model in the pool):

- A 120B-class open-weight model
- Mistral Large 3 (675B parameter class) [14]
- A 27B-class open-weight model
- Amazon Nova Pro v1 [1]
- DeepSeek V3.2 [4]
- Mistral Magistral Small 2509 [13]

The three verification judges are Amazon Nova Pro [1], Mistral Large 2 [12], and a 70B-class model from a third provider used only for verification.

C Per-Category Breakdown

Table 7 reveals that generation strategy affects categories differently. The largest variation (Δmax = 21.5 pp) occurs in C3 (classification), where S3 scores 87.5% but S4 only 66.0%; hybrid injection into real documents makes classification decisions substantially harder. C1 (LLM compensation) shows 19.3 pp variation, with S1 and S5 scoring 80.9% but S1-lite only 61.6%. C6 (multi-author) shows 16.7 pp variation (S0 highest at 57.8, S4 lowest at 41.1). C2 (retrieval bottleneck) is most stable across strategies (Δmax = 10.8 pp), suggesting that retrieval difficulty is relatively robust to generation method, an expected result since retrieval depends on embedding similarity rather than document surface properties. We also examined whether these per-category differences reorder the categories (Kendall's τ between strategy pairs), as a proxy for effects on system ranking. With only 6 categories the rank correlations are not distinguishable from chance under a permutation test (all $p > 0.05$).