PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 31-47
E-ISSN: 3109-8886

Confidence-Gated Long-Term Memory for Multi-Session LLM Assistants with Temporal Retrieval, Conflict Revision, and Selective Forgetting

Samuel Thompson[1], Ming Xu[2], Olivia Brooks[3]

[1]Departement of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA
[2]Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[3]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

Email: [1*]s.thompson@gmail.com

Abstract

Long-lived language-model assistants must decide what to store, which fact version to retrieve, when old evidence remains relevant, and when to reject an unsupported premise. This study evaluates CG-TRF, a confidence-gated temporal retrieval and forgetting policy, on the final ten-conversation LoCoMo benchmark: 272 sessions, 5,882 turns, 2,541 atomic observations, and 1,986 multi-hop, temporal, open-domain, single-hop, or adversarial questions. A fixed CPU-only extractive probe compared full history, a 120-turn window, session summaries, word TF-IDF, word-character hybrid memory, temporal-decay retrieval, and CG-TRF. CG-TRF combined confidence-gated writes, temporal and speaker alignment, revision links, 70% retention, diverse top-five retrieval, and confidence-based premise rejection. On 1,540 answer-bearing questions, it attained 0.1375 normalized token F1 and 0.0227 exact match, versus 0.1364 and 0.0175 for unrestricted hybrid memory. That F1 difference was not statistically supported, although CG-TRF stored 30% fewer records. It rejected 50.22% of 446 adversarial questions and reduced tempting-answer overlap to 0.0381; nonselective baselines rejected 0%. Its primary-budget Recall@5 was 0.4275 versus 0.5047 for hybrid memory; without retention, Recall@5 reached 0.5163 and F1 reached 0.1463. Confidence and attribution improved memory-layer safeguards, but forgetting caused the principal retrieval loss.

Keywords: *Large Language Models, Long-Term Conversational Memory, Multi-Session Dialogue, Temporal Retrieval, Retrieval-Augmented Generation, Selective Prediction, Memory Revision, Locomo.*

Abstrak

Asisten berbasis model bahasa jangka panjang harus memutuskan informasi apa yang akan disimpan, versi fakta mana yang akan diambil, kapan bukti lama masih relevan, dan kapan harus menolak premis yang tidak didukung. Studi ini mengevaluasi CG-TRF—sebuah kebijakan pengambilan dan pelupaan temporal dengan mekanisme *confidence-gated* (penyaringan berbasis tingkat keyakinan)—pada tolok ukur LoCoMo, yang mencakup sepuluh percakapan panjang: 272 sesi, 5.882 giliran bicara (*turns*), 2.541 observasi atomik, dan 1.986 pertanyaan yang mencakup kategori *multi-hop*, temporal, domain terbuka, *single-hop*, maupun adversarial. Mekanisme *probe* ekstraktif berbasis CPU digunakan untuk membandingkan berbagai metode: riwayat lengkap, jendela 120 giliran bicara, ringkasan sesi, TF-IDF kata, memori hibrida kata-karakter, pengambilan dengan peluruhan temporal, dan CG-TRF. CG-TRF menggabungkan penulisan dengan penyaringan keyakinan (*confidence-gated writing*), penyelarasan temporal dan pembicara, penautan revisi, retensi 70%, pengambilan lima hasil teratas yang beragam, serta penolakan premis berbasis keyakinan. Untuk 1.540 pertanyaan yang dapat dijawab, metode ini mencapai skor F1 token ternormalisasi sebesar 0,1375 dan skor kecocokan eksak (*exact match*) sebesar 0,0227, dibandingkan dengan 0,1364 dan 0,0175 untuk memori hibrida tanpa batasan. Perbedaan skor F1 tersebut tidak signifikan secara statistik, meskipun CG-TRF menyimpan catatan 30% lebih sedikit. Metode ini menolak 50,22% dari 446 pertanyaan adversarial dan mengurangi tingkat tumpang tindih jawaban yang menjebak (*tempting-answer overlap*) menjadi 0,0381, sedangkan metode dasar non-selektif menolak 0%. Nilai Recall@5 dengan anggaran utama adalah 0,4275, dibandingkan dengan 0,5047 untuk memori hibrida. Tanpa retensi, Recall@5 mencapai 0,5163 dan F1 mencapai 0,1463. Mekanisme keyakinan dan atribusi meningkatkan perlindungan lapisan memori; namun, proses pelupaan merupakan penyebab utama penurunan kinerja pengambilan informasi.

Kata Kunci: Model Bahasa Besar, Memori Percakapan Jangka Panjang, Dialog Multi-sesi, Pengambilan Temporal, *Retrieval-Augmented Generation*, Prediksi Selektif, Revisi Memori, LoCoMo.

A. INTRODUCTION

An assistant operating across months must preserve commitments, retrieve cross-session evidence, reconcile

Diterima Redaksi: 01-08-2026 | Selesai Revisi: 06-08-2026 | Diterbitkan Online: 07-08-2026

updates, and reject wrong-person premises. LoCoMo tests these demands in ten long, human-refined agent conversations with QA, evidence identifiers, generated observations and summaries, and annotated events [1], [2]. It covers factual, multi-session, temporal, commonsense, and adversarial questions; tested long-context and retrieval systems remained below human performance, especially on time [1].

Persistent memory is not nominal context length. LongMemEval evaluates extraction, temporal reasoning, updating, and abstention [3]; Multi-Session Chat benefits from retrieved utterances and summaries [4]; LongBench treats long context as multitask [5]; and relevant material can remain underused inside the window [6]. Full transcripts are therefore availability mechanisms, not complete memory policies.

Architectures separate memory functions. MemGPT frames context as virtual memory [7]; MemoryBank combines storage, update, recall, and forgetting [8]; LongMem decouples retrieval [9]; generative agents rank relevance, recency, and importance [10]; and reflective or segmented memories vary granularity [11], [12]. One answer score can conceal failures in admission, representation, retrieval, or retirement.

RAG connects external memory to an LLM [13]. BM25 remains a lexical reference [14], while Sentence-BERT, dense passage retrieval, and ColBERT add learned semantic or late-interaction scoring [15]–[17], and reciprocal-rank fusion combines rankings [18]. Recurrent and nearest-neighbor memories extend dependencies beyond fixed segments [19], [20]. This study uses CPU-reproducible word and character TF-IDF to isolate policy effects.

Time changes relevance. Temporal commonsense [21], time-conditioned QA [22], streaming evidence [23], and temporal knowledge-graph QA [24] all expose this difficulty. Historical questions require old evidence, whereas current-state questions should demote superseded values; temporal graph memory can preserve both views [25].

Answer policy also needs uncertainty. Calibration relates confidence to correctness [26], selective prediction declines low-confidence cases [27], and conformal methods provide coverage under stated assumptions [28]. CG-TRF instead uses a transparent retrieval-and-attribution gate and makes no neural-calibration or distribution-free guarantee.

Applied systems reinforce modularity. TimeRAG-Memory gates writes and supports recency-aware overwrite, decay, and deletion [29]; privacy-aware topic preferences provide compact routing signals [30]; and retrieved behavior or sessions support recommendation and counseling [31], [32]. Review-grounded recommendation tests textual support [33], while strategy-controlled emotional support retrieves human responses [34].

Retrieval coverage alone does not establish answer support. Word-level attribution [35], deterministic evidence chains [36], evidence-calibrated abstention [37], and budgeted multi-hop retrieval [38] motivate separate provenance, refusal, and context-cost measures. Harmful-behavior refusal [39], evidence cards [40], self-verification [41], data-integrity oversight [42], and multilingual trust-calibrated RAG [43] broaden deployment evaluation. CG-TRF addresses only wrong-person premises—not jailbreaks, toxicity, prompt injection, or multilingual safety.

The question is whether a lightweight policy can reduce context and false attribution without hiding retrieval loss. Unlike TimeRAG-Memory's end-to-end dialogue evaluation [29], we audit all LoCoMo-10 questions with one fixed probe, speaker-specific adversarial tests, clustered inference, and a storage frontier. CG-TRF scores official observations, links value changes, protects current linked records, and combines lexical, character, temporal, speaker, and diversity signals. Contributions are the normalized audit; seven-condition answer, evidence, attribution, cost, and latency evaluation; a 25%–100% forgetting frontier; distance and component analyses; and conversation-clustered inference..

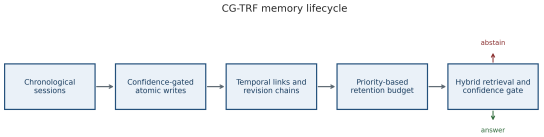


Figure 1 CG-TRF Memory Lifecycle. Chronological Sessions Are Converted To Atomic Candidate Records, Screened For Informativeness, Linked Across Time, Retained Under A Budget, And Retrieved Subject To An Answer-Confidence Gate

B. METODE

Dataset, Scope, And Audit

The complete locomo10.txt contained 10 conversations (January 2022–January 2024), 272 list-valued sessions, 5,882 turns, 133,772 whitespace words, 272 generated summaries, 2,541 generated observations, 669 annotated events, and 1,986 QA records. Sixteen orphan timestamp keys were excluded. Turn text included available BLIP captions; unavailable image binaries and metadata were excluded. Annotated events were never retrieval input, preserving the release's generated-memory versus annotated-evaluation distinction [1], [2].

Table 1 Locomo-10 Composition By Conversation

Conversation	Sess.	Turns	Obs.	Summary sent.	QA	Study tokens
conv-26	19	419	184	208	199	15042
conv-30	19	369	169	126	105	11548
conv-41	32	663	324	226	193	22509
conv-42	29	629	266	197	260	18646
conv-43	29	680	267	184	242	22283

Conversation	Sess.	Turns	Obs.	Summary sent.	QA	Study tokens
conv-44	28	675	277	194	158	21437
conv-47	31	689	268	208	190	20373
conv-48	30	681	291	203	239	18872
conv-49	25	509	240	161	196	16337
conv-50	30	568	255	214	204	20812

Categories 1–4 supplied 1,540 answer-bearing items; four without evidence remained in QA but not the 1,536-item evidence denominator. Category 5 supplied 446 premise challenges: 444 had a tempting adversarial_answer, and two answered "No." Primary rejection included all 446; a field-faithful variant was a sensitivity check.

Nine malformed evidence fields were normalized by splitting compounds, repairing delimiters or leading characters, deduplicating, and replacing two nonexistent identifiers after unambiguous localization (D10:19→D20:15; D4:36→D13:3). The same validated parser flattened fifteen multi-source observations. Primary results retained eleven duplicate questions and one label conflict; a 1,974-item sensitivity set removed them. Official temporal answers remained primary, with two literal-weekday corrections tested separately.

Table 2 Evaluation Sets And Audit Treatment

Audit item	Count	Treatment
Official QA records	1986	Retained in primary evaluation
Answer-bearing categories 1-4	1540	QA F1 and exact match
Evidence-bearing answer QA	1536	Retrieval metrics
Adversarial questions	446	Premise-rejection evaluation
Strict positive/adversarial pairs	428	Paired attribution analysis
Formatting-anomalous evidence fields	9	Deterministically normalized
Duplicate or ambiguous rows	12	Retained primary; removed in clean sensitivity

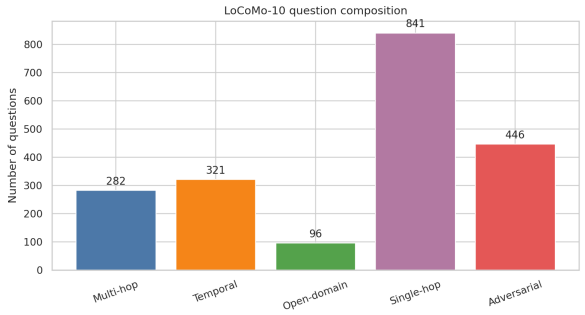


Figure 2 Distribution Of The 1,986 Locomo Questions Across Reasoning Categories

Memory Representations And Comparison Conditions

Every method received the same questions and chronology; none saw gold answers, labels, evidence, or adversarial answers. Word TF-IDF used lowercase alphanumeric 1–2-grams, fixed stop words, sublinear frequency, and L2 normalization; character TF-IDF used boundary-aware 3–5-grams. A fixed regex counted study-context tokens, not billable API tokens.

Full history scanned all raw turns and charged the entire history; a window scanned the latest 120 turns. Summary memory retrieved five sentences by word-character score and used coarse session-level evidence matching. Vector memory indexed released observations with word TF-IDF; hybrid used 72% word and 28% character cosine; temporal-decay RAG added 16% exponential session recency. These RAG baselines [13] are simpler than learned retrievers [15]–[17]. CG-TRF used the same 2,541 generated, speaker-explicit observations [2], retaining 70% per conversation. Raw-turn comparisons mix representation and policy; observation-based conditions isolate policy effects.

Applied retrieval studies likewise separate stages: private-runbook QA [44], code retrieval plus summaries [45], FiQA reranking [46], and retrieval-quality prediction [47] distinguish retrieval from final response quality. Rank-aggregation theory supplies broader fusion context [48]. These are methodological analogues, not evidence about conversational memory.

Table 3 Memory Methods, Representations, Access Policies, And Primary Context Behavior

Method	Memory unit	Active store	Query policy
Full history	Raw turns	All turns	Lexical scan; full history cost
Sliding window	Raw turns	Last 120 turns	Lexical scan
Session summaries	Summary sentences	All summaries	Word + character TF-IDF
Vector memory	Observation facts	All facts	Word TF-IDF
Hybrid vector	Observation facts	All facts	72% word + 28% character
Temporal-decay RAG	Observation facts	All facts	Word score + recency
CG-TRF	Observation facts	70% priority budget	Hybrid + time + speaker + MMR + reject

Confidence-Gated Writes And Selective Forgetting

Without QA labels, write confidence weighted length saturation, explicit speaker, date/number, action verbs, and novelty by 0.30/0.20/0.20/0.15/0.15. Novelty was one minus maximum prior same-speaker word-TF-IDF cosine; scores ≥0.38 were eligible, with fallback to all records if required by the budget.

Retention priority used 52% write confidence, 23% novelty, 15% normalized recency, plus 0.25 for the latest revision-chain record. Per-conversation targets were 70% primarily and 25%, 40%, 55%, 85%, or 100% in sensitivity

tests. Record-count budgets need not shrink the five returned records' token total.

This is active-store eviction, not deletion compliance: LoCoMo lacks deletion requests or authoritative erasure. It follows importance/recency retention [8], [10] and explicit tiers [7], but—unlike user-requested deletion [29] or private preference aggregation [30]—claims neither cryptographic erasure nor differential privacy.

Temporal Retrieval And Conflict-Linked Records

Retrieval combined word and character cosine. If a question named one participant, that person's records gained 0.105 and the other's lost 0.135. Temporal weight was ≤0.16: explicit dates used 60-day exponential proximity; current-state cues used session recency; other cues rewarded temporal expressions without globally penalizing old evidence.

Revision proxies linked same-speaker, cross-session records at word-TF-IDF similarity ≥0.34 when dates, numbers, statuses, or negation differed. The newest link was active; older links lost 0.16 only for current-state questions, while historical questions retained access. This detects near-duplicate value change, not annotated contradiction, following the active/history distinction [25]. Top-five maximal marginal relevance reranked forty candidates with 86% relevance and 14% diversity.

Adjacent systems illustrate why structured evidence channels matter: accounting-aware retrieval routes evidence types [49], multi-source incident analysis separates attribution streams [50], and narrative-aware claim verification studies auxiliary ranking cues [51]. They motivate, but do not validate, CG-TRF's time and speaker features.

Answer Extraction And Confidence-Based Rejection

A fixed probe chose the clause with greatest content-word overlap, resolved relative dates, and applied question-type extraction for counts, people, places, causes, instruments, and polarity; otherwise it removed a leading subject/reporting verb and returned ≤24 words. It tests access to answer-bearing language, not foundation-model generation.

Only CG-TRF abstained for absent records, weak top score/margin, or wrong-speaker evidence. If the unconstrained lexical winner exceeded the named person's best record by ≥0.08, it returned “I don't know.” This implements selective risk [27] without calibration [26] or conformal-coverage [28] claims. It is evidence-sufficiency abstention [37], not harmful-behavior filtering [39]. Retrieval-grounded log narratives [52], prototype explanations [53], regulation retrieval [54], and command-safety evaluation [55] provide cross-domain examples of separating evidence, explanation, and action policy.

Other domains operationalize rejection and uncertainty with risk–coverage analysis [56], calibrated/conformal governance [57], cross-dataset selective refusal [58], and calibration-aware late fusion [59]. They support separate evaluation of ranking, calibration, and review—not portability of their thresholds to LoCoMo.

Metrics And Statistical Analysis

Normalized F1 lowercased and removed punctuation/articles before token precision/recall; exact match used the same normalization. Evidence metrics were Recall@5, Hit@5, complete-set All-Evidence@5, MRR, and nDCG@5. Summaries used session-level evidence; others used turns. Safety metrics were category-5 abstention and overlap with the tempting answer (lower is safer).

Fixed cues defined current-state cases. Conflict revision required the latest gold session to outrank an earlier non-gold candidate; stale citation meant an older non-gold top record. Only 22 cases per method qualified. Strict attribution paired 428 adversarial items with answerable items sharing evidence and the tempting/gold answer; success required positive evidence plus adversarial rejection, with an extractive variant also requiring nonzero positive F1.

With seed 20260804, paired uncertainty used 10,000 conversation-cluster bootstrap resamples across ten conversations. Exact two-sided sign-flip tests enumerated 1,024 assignments with within-metric Holm adjustment. Post-index CPU wall time is implementation-specific because scan volumes differ.

C. RESULTS AND DISCUSSION

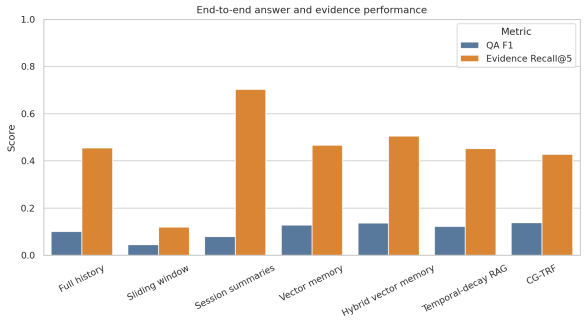
Overall Answer Recovery And Evidence Retrieval

CG-TRF's primary-budget F1 was 0.1375, versus 0.1364 for hybrid and 0.1282 for vector memory; exact match was 0.0227. Its hybrid F1 difference was 0.0022 (clustered 95% interval, -0.0091–0.0151; adjusted p = 0.7539), an empirical tie. Versus full history, the gain was 0.0363 (0.0209–0.0517; adjusted p = 0.0234), favoring atomic controlled retrieval for this probe.

Table 4 Overall Answer, Retrieval, Safety, And Context Results. Scores Are Proportions

Method	QA F1	EM	Ev. R@5	All ev.	Adv. reject	Wrong-answer overlap	Ctx. tokens
Full history	0.1011	0.0260	0.4554	0.4186	0.0000	0.0855	19,126.10
Sliding window	0.0449	0.0123	0.1196	0.1074	0.0000	0.0422	3,786.11
Session summaries	0.0789	0.0117	0.7023	0.6445	0.0000	0.0904	3,369.06
Vector memory	0.1282	0.0162	0.4670	0.4245	0.0000	0.1414	79.29
Hybrid vector memory	0.1364	0.0175	0.5047	0.4577	0.0000	0.1485	82.34

Method	QA F1	EM	Ev. R@5	All ev.	Adv. reject	Wrong-answer overlap	Ctx. tokens
Temporal-decay RAG	0.1220	0.0149	0.4518	0.4102	0.0000	0.1372	82.91
CG-TRF	0.1375	0.0227	0.4275	0.3815	0.5022	0.0381	88.15



Gambar 1 QA F1 And Evidence Recall@5 For All Seven Memory Conditions

At 70% retention, CG-TRF Recall@5 was 0.4275, below full history (0.4554), vector (0.4670), and hybrid (0.5047); All-Evidence@5 was 0.3815. Summaries reached 0.7023/0.6445, but session-level matching does not prove detail preservation. At 100%, CG-TRF reached 0.5163/0.4661, slightly above hybrid; eviction, not its full-store ranking, caused the main recall loss.

Summary memory led coarse recall in every category. Among turn-linked observations, hybrid led multi-hop/single-hop/temporal recall (0.2362/0.5777/0.6285) versus CG-TRF (0.1972/0.4909/0.5350). CG-TRF nevertheless had slightly higher multi-hop F1 (0.0980 vs. 0.0953) and temporal F1 of 0.1521, between full history (0.1618) and hybrid (0.1189). Open-domain extraction remained weak.

Table 5 Evidence Recall@5 By Memory Method And Question Category

Method	Multi-hop	Open-domain	Single-hop	Temporal
CG-TRF	0.1972	0.1783	0.4909	0.5350
Full history	0.1587	0.2328	0.5349	0.5714
Hybrid vector memory	0.2362	0.2292	0.5777	0.6285
Session summaries	0.4393	0.4540	0.7818	0.7965
Sliding window	0.0421	0.0480	0.1427	0.1480
Temporal-decay RAG	0.1946	0.1785	0.5135	0.5942
Vector memory	0.2067	0.1728	0.5349	0.6020

Table 6 Normalized Token F1 By Memory Method And Answer-Bearing Category

Method	Multi-hop	Open-domain	Single-hop	Temporal
CG-TRF	0.0980	0.0812	0.1516	0.1521
Full history	0.0499	0.0972	0.0956	0.1618
Hybrid vector memory	0.0953	0.0761	0.1638	0.1189
Session summaries	0.0662	0.1151	0.0921	0.0445
Sliding window	0.0362	0.1118	0.0396	0.0465
Temporal-decay RAG	0.0903	0.0728	0.1490	0.0939
Vector memory	0.0933	0.0739	0.1553	0.1042

Attribution Conflict, Current-State Diagnostics, And Abstention

Nonselective baselines abstained 0%. CG-TRF rejected 0.5022 of 446 adversarial items and reduced tempting-answer overlap to 0.0381, versus 0.1485 hybrid, 0.1414 vector, and 0.0855 full history. Removing the speaker gate cut abstention to 0.0202 with nearly unchanged F1. On 428 strict positive/adversarial pairs, evidence-attribution and extractive-pair success were 0.2991 and 0.3458; always-answer baselines scored zero under paired rejection.

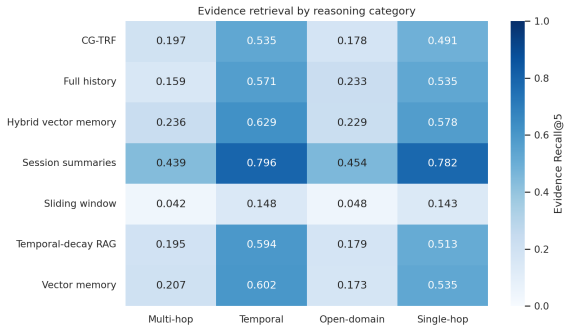


Figure 3 Evidence Recall@5 By Reasoning Category

Attribution conflict, current-state diagnostics, and abstention
Nonselective baselines abstained 0%. CG-TRF rejected 0.5022 of 446 adversarial items and reduced tempting-answer overlap to 0.0381, versus 0.1485 hybrid, 0.1414 vector, and 0.0855 full history. Removing the speaker gate cut abstention to 0.0202 with nearly unchanged F1. On 428 strict positive/adversarial pairs, evidence-attribution and extractive-pair success were 0.2991 and 0.3458; always-answer baselines scored zero under paired rejection.

Table 7 Adversarial Rejection And Current-State Retrieval Diagnostics

Method	Adv. reject	Wrong-answer overlap	Current-state hit	Stale top citation
Full history	0.0000	0.0855	0.2727	0.2273
Sliding window	0.0000	0.0422	0.0909	0.0909
Session summaries	0.0000	0.0904	0.4545	0.2727
Vector memory	0.0000	0.1414	0.4545	0.3636

Method	Adv. reject	Wrong-answer overlap	Current-state hit	Stale top citation
Hybrid vector memory	0.0000	0.1485	0.5000	0.3636
Temporal-decay RAG	0.0000	0.1372	0.4545	0.1818
CG-TRF	0.5022	0.0381	0.5000	0.1818

Half remained unrejected, and speaker mismatch is benchmark-specific because one person may describe another. The rule is an attribution safeguard, not universal answerability. Evolving jailbreak defenses address a different attack-distribution problem [60]; numerical consistency validators likewise check properties beyond retrieval support [61].

On the small current-state subset, CG-TRF's revision/stale-citation rates were 0.5000/0.1818. Hybrid was 0.5000/0.3636; temporal-decay RAG also had 0.1818 staleness. Removing the revision penalty changed no displayed aggregate score because chains were sparse. These are version-link diagnostics, not validated contradiction resolution.

Selective Forgetting And Long-Range Robustness

The retention curve was monotonic: F1/Recall@5 rose from 0.0957/0.2071 at 25% through 0.1375/0.4275 at 70% to 0.1463/0.5163 at 100%. The final 30 storage points added 0.0088 F1 and 0.0888 recall; evidence suffers more because partial answer words can survive missing multi-evidence members.

Table 8 CG-TRF Performance Under Six Record-Retention Budgets

Record budget	QA F1	Ev. R@5	All ev.	Active records	Ctx. tokens
25%	0.0957	0.2071	0.1810	64.95	94.94
40%	0.1146	0.2976	0.2617	103.88	91.53
55%	0.1287	0.3710	0.3307	142.91	88.58
70%	0.1375	0.4275	0.3815	181.48	88.15
85%	0.1414	0.4808	0.4342	220.40	86.64
100%	0.1463	0.5163	0.4661	258.79	83.62

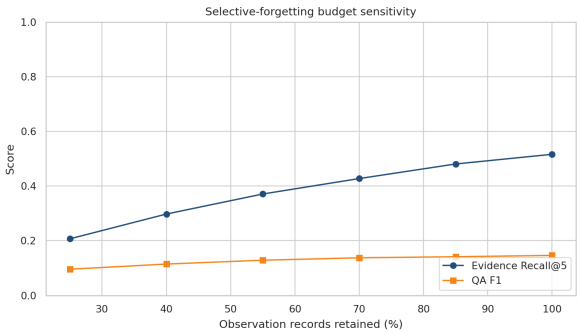


Figure 4 QA F1 And Evidence Recall@5 As A Function Of The Retained-Record Budget

Context was nonmonotonic: 25% retention averaged 94.94 tokens versus 83.62 at 100%, because priority favored denser records and retrieval returned five units. Record budgets shrink the index, not necessarily prompts. All compact retrievers worsened with evidence distance, especially the window. CG-TRF avoided global recency decay but evicted facts without future-query knowledge, so write gates need durable-utility signals beyond surface informativeness [3].

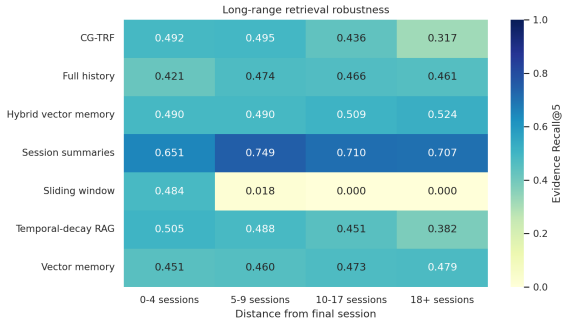


Figure 5 Evidence Recall@5 By Distance Between Gold Evidence And The Final Conversation Session

Component Contributions, Efficiency, And Selective Risk

Without temporal scoring, F1 fell from 0.1375 to 0.1288 while Recall@5 rose from 0.4275 to 0.4316: time aided answer selection but sometimes displaced literal evidence, paralleling auxiliary-cue tradeoffs in claim verification [51]. Word-only retrieval reduced F1/recall to 0.1310/0.3983. Removing speaker ranking lowered MRR from 0.3943 to 0.3746; disabling speaker rejection collapsed abstention. Revision had no displayed effect; removing forgetting improved F1/recall and enlarged the mean index from 181.48 to 258.79.

Table 9 CG-TRF Component Ablation And Unrestricted-Memory Sensitivity

Variant	QA F1	Ev. R@5	MRR	Adv. reject	Stale top	Active records
Full CG-TRF	0.1375	0.4275	0.3943	0.5022	0.1818	181.48
Without temporal scoring	0.1288	0.4316	0.3942	0.5022	0.1818	181.48
Without speaker gate	0.1376	0.4170	0.3746	0.0202	0.1818	181.48
Without revision penalty	0.1375	0.4275	0.3943	0.5022	0.1818	181.48
Word-only retrieval	0.1310	0.3983	0.3691	0.5022	0.1818	181.48
Without forgetting (100%)	0.1463	0.5163	0.4621	0.5000	0.2273	258.79

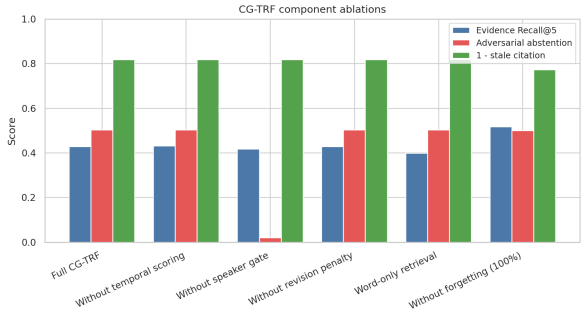


Figure 6 Effects Of Temporal, Speaker, Revision, Lexical, And Forgetting Components

Full history exposed 19,126.10 study tokens per answer-bearing question, versus 79.29–88.15 for compact top-five methods (>99.5% less). CG-TRF median/p95 latency was 6.71/22.51 ms versus hybrid's 1.19/9.29 ms. Million-record stores would need approximate, batched, or cost-aware routing [62]; dense or late-interaction retrieval [16]–[18] would add dependencies.

Table 10 Active Records, Returned-Context Tokens, And Measured Query Latency

Method	Ctx. tokens	Median ms	P95 ms	Active records
Full history	19,126.10	5.6498	20.3974	601.28
Sliding window	3,786.11	0.4840	8.4838	120.00
Session summaries	3,369.06	1.7142	9.6214	195.31
Vector memory	79.29	0.4745	4.6729	258.79
Hybrid vector memory	82.34	1.1870	9.2911	258.79
Temporal-decay RAG	82.91	0.0916	0.1807	258.79
CG-TRF	88.15	6.7131	22.5116	181.48

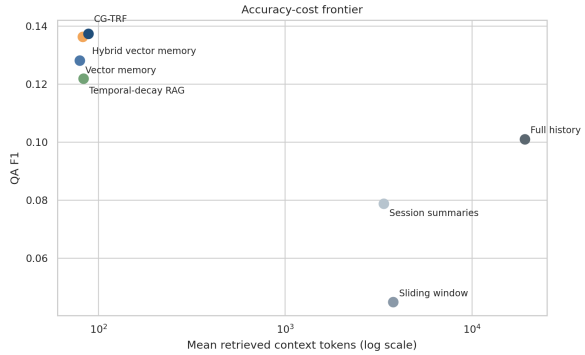


Figure 7 QA F1 Versus Mean Returned-Context Tokens On A Logarithmic X-Axis

F1 declined from the highest-confidence decile toward full coverage. This shows ordering, not calibrated probability: deployment needs independent conversations and a stable target [26], and ten groups cannot justify strong conformal claims [28]. Evidence-calibrated QA [37], multilingual RAG [43], and trajectory-level reliability routing [63] reinforce separate answerability, calibration, and group-split evaluation.

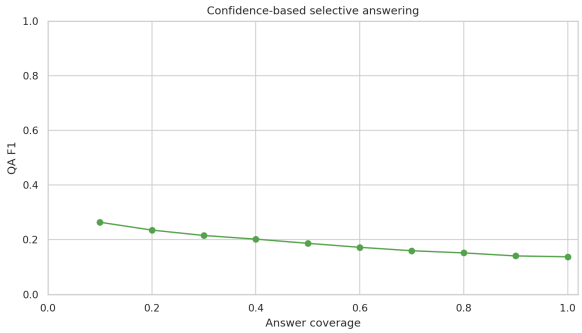


Figure 8 QA F1 As Answer Coverage Increases From 10% To 100% Under CG-TRF Confidence Ranking

Statistical Interpretation, Data Sensitivity, And Limitations

Clustered inference qualified small differences. F1 gains over summaries, window, and full history excluded zero (adjusted p = 0.0117, 0.0117, 0.0234); differences from hybrid and vector were unsupported. Recall was worse than hybrid and summaries but better than the window, confirming stage-specific metrics.

Table 11 Selected Conversation-Clustered Differences Between CG-TRF And Baselines

Metric	Comparison	Difference	Cluster 95% CI	Holm p
qa f1	vs. Full history	0.0363	[0.0209, 0.0517]	0.0234
qa f1	vs. Hybrid vector memory	0.0022	[-0.0091, 0.0151]	0.7539
qa f1	vs. Temporal-decay RAG	0.0154	[0.0038, 0.0285]	0.1582
evidence recall	vs. Full history	-0.0270	[-0.0578, -0.0011]	0.1113
evidence recall	vs. Hybrid vector memory	-0.0761	[-0.0916, -0.0607]	0.0117
evidence recall	vs. Sliding window	0.3057	[0.2737, 0.3347]	0.0117

Removing eleven duplicates and one label conflict changed no qualitative ranking; two weekday corrections barely changed temporal F1. Still, LoCoMo uses human-refined agent conversations [1] and pre-generated memories [2], so names, repetition, and artifact structure may matter. Chronological drift requires revalidation under changing regimes [64], while held-out-user transfer should be tested rather than assumed [65].

The fixed probe isolates retrieval but underestimates inference and cannot measure hallucination. Summary evidence is coarse; conflict chains are proxies; the current-state subset is small; ten conversations limit power; and latency is implementation-specific. Eviction supplies neither secure deletion, consent, correction, nor privacy.

Future grouped studies should cross-fit write and answerability gates, test correction/deletion, and equalize token budgets. Generators should preserve retrieval, attribution, and hallucination metrics [35], [36] and expose multi-hop cost [38]. Review interfaces could combine

existing evidence-card patterns [40], [42] with linked counseling recommendations [66], trust-calibrated product cards [67], accounting evidence cards [68], incident cards [69], and text-grounded rationale cards [70]. These adjacent designs motivate visible sources, confidence, risk, and human override—not their domain-specific conclusions. Operational analogues separate conformal alert thresholds from evidence tickets [71], preserve provenance in incident triage [72], and localize forensic narratives to detected windows [73]. Recurrent, lookup, temporal, and segmented memories remain complementary [11], [12], [19]–[24].

D. CONCLUSIONS

Across 1,986 questions, seven conditions, six budgets, ablations, distance strata, adversarial pairs, and clustered uncertainty, CG-TRF returned about 88 study tokens instead of 19,126. It matched hybrid-memory F1 while storing 30% fewer observations and rejected about half of wrong-person premises.

That saving was not free: 70%-budget Recall@5 was 0.4275 versus hybrid's 0.5047; full retention raised CG-TRF to 0.5163 recall and 0.1463 F1. Revision effects were inconclusive. Assistants should preserve source/time metadata, gate attribution conflicts, retain historical versions, and measure eviction loss; source-grounded incident summaries offer a parallel provenance pattern [72]. Confidence should govern answering without concealing forgotten evidence.

E. DAFTAR PUSTAKA

[1] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, “Evaluating Very Long-Term Conversational Memory of LLM Agents,” in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers, 2024, pp. 13851–13870, doi: 10.18653/v1/2024.acl-long.747.

[2] Snap Research, “LoCoMo: Data and Code for Evaluating Very Long-Term Conversational Memory of LLM Agents,” GitHub, 2024. [Online]. Available: https://github.com/snap-research/locomo. Accessed: Aug. 4, 2026.

[3] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu, “LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory,” in Proc. International Conference on Learning Representations (ICLR), 2025.

[4] J. Xu, A. Szlam, and J. Weston, “Beyond Goldfish Memory: Long-Term Open-Domain Conversation,” in Proc. 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 5180–5197, doi: 10.18653/v1/2022.acl-long.356.

[5] Y. Bai et al., “LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding,” in Proc. 62nd Annual Meeting of the Association for Computational Linguistics, 2024, pp. 3119–3137, doi: 10.18653/v1/2024.acl-long.172.

[6] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, “Lost in the Middle: How Language Models Use Long Contexts,” Transactions of the Association for Computational Linguistics, vol. 12, pp. 157–173, 2024, doi: 10.1162/tacl_a_00638.

[7] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, “MemGPT: Towards LLMs as Operating Systems,” arXiv:2310.08560, 2023.

[8] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, “MemoryBank: Enhancing Large Language Models with Long-Term Memory,” in Proc. AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp. 19724–19731, 2024, doi: 10.1609/aaai.v38i17.29946.

[9] W. Wang, L. Dong, H. Cheng, X. Liu, X. Yan, J. Gao, and F. Wei, “Augmenting Language Models with Long-Term Memory,” in Advances in Neural Information Processing Systems, vol. 36, pp. 74530–74543, 2023.

[10] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative Agents: Interactive Simulacra of Human Behavior,” in Proc. 36th Annual ACM Symposium on User Interface Software and Technology, Art. 2, pp. 1–22, 2023, doi: 10.1145/3586183.3606763.

[11] Z. Tan et al., “In Prospect and Retrospect: Reflective Memory Management for Long-Term Personalized Dialogue Agents,” in Proc. 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 8416–8439, doi: 10.18653/v1/2025.acl-long.413.

[12] Z. Pan et al., “SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents,” in Proc. International Conference on Learning Representations, 2025.

[13] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[14] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[15] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

[16] V. Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering,” in Proc. EMNLP, 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.

[17] O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” in Proc. 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 39–48, doi: 10.1145/3397271.3401075.

[18] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in Proc. 32nd International ACM SIGIR Conference, 2009, pp. 758–759, doi: 10.1145/1571941.1572114.

[19] [19] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," in Proc. 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2978–2988, doi: 10.18653/v1/P19-1285.

[20] Y. Wu, M. N. Rabe, D. Hutchins, and C. Szegedy, "Memorizing Transformers," in Proc. International Conference on Learning Representations, 2022.

[21] L. Qin, A. Gupta, S. Upadhyay, L. He, Y. Choi, and M. Faruqui, "TimeDial: Temporal Commonsense Reasoning in Dialog," in Proc. ACL-IJCNLP, 2021, pp. 7066–7076, doi: 10.18653/v1/2021.acl-long.549.

[22] M. Zhang and E. Choi, "SituatedQA: Incorporating Extra-Linguistic Contexts into QA," in Proc. EMNLP, 2021, pp. 7371–7387, doi: 10.18653/v1/2021.emnlp-main.586.

[23] A. Liška et al., "StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models," in Proc. 39th International Conference on Machine Learning, PMLR, vol. 162, pp. 13604–13622, 2022.

[24] A. Saxena, S. Chakrabarti, and P. Talukdar, "Question Answering Over Temporal Knowledge Graphs," in Proc. ACL-IJCNLP, 2021, pp. 6663–6676, doi: 10.18653/v1/2021.acl-long.520.

[25] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef, "Zep: A Temporal Knowledge Graph Architecture for Agent Memory," arXiv:2501.13956, 2025.

[26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. 34th International Conference on Machine Learning, PMLR, vol. 70, pp. 1321–1330, 2017.

[27] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Advances in Neural Information Processing Systems, vol. 30, pp. 4878–4887, 2017.

[28] A. N. Angelopoulos and S. Bates, "Conformal Prediction: A Gentle Introduction," Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.

[29] X. Sun, Y. Lu, and J. Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting," J. Adv. Comput. Syst., vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[30] M.-J. Kuo, D. Zheng, and J. Hires, "Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385–401, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[31] Q. Xin, "Behavior Retrieval plus Response Generation for Interpretable Conversational Personalized Recommendation," IJEEPSE, vol. 9, no. 2, pp. 120–136, 2026, doi: 10.31258/ijeepse.9.2.120-136.

[32] Y. Zhang and H. Zhang, "A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464–486, Aug. 2025, doi: 10.51903/jtie.v4i2.547.

[33] X. Chang, Y. Lu, and Z. S. Zhong, "Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 9–22, 2026, doi: 10.54732/jeecs.v11i1.2.

[34] [34] Y. Zhang and Z. Zhou, "Strategy-Aware Therapist Imitation for Emotional Support Dialogues: A Reproducible ESConv Study for LLM Response Control," Adv. Educ. Technol. Psychol., vol. 10, no. 2, pp. 92–97, 2026, doi: 10.23977/aetp.2026.100213.

[35] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76–92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[36] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520–533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[37] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502–520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[38] W. Su, S. Chen, and C. Zhao, "Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649–662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[39] D. Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[40] [40] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397–414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[41] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification," J.

Adv. Comput. Syst., vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[42] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186–191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[43] Y. Chen and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141–164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[44] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109–125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[45] Y. Li, "Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," J. Adv. Comput. Syst., vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[46] S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361–378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[47] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality Analysis and Evaluation Prediction of RAG Retrieval Based on Machine Learning Algorithms," arXiv:2511.19481, 2025, doi: 10.48550/arXiv.2511.19481.

[48] Z. S. Zhong and S. Ling, "Improved Theoretical Guarantee for Rank Aggregation via Spectral Method," Inf. Inference: J. IMA, vol. 13, no. 3, Art. iaae020, Sep. 2024, doi: 10.1093/imaiai/iaae020.

[49] Y. Chen, S. Zhou, and E. Lin, "Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649–663, Dec. 2025, doi: 10.51903/jtie.v4i3.542.

[50] G. Liu, S. He, and I. Liu, "LLM-Augmented Multi-Source Root Cause Attribution for CPU and Network Faults in Microservices," J. Adv. Comput. Syst., vol. 3, no. 6, pp. 39–57, Jun. 2023, doi: 10.69987/JACS.2023.30604.

[51] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327–340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[52] X. Sun, Z. S. Zhong, and Q. Wu, "Retrieval-Grounded HDFS Log Anomaly Detection and Deterministic Failure Narrative Generation," J. Comput. Syst. Appl., vol. 3, no. 1, pp. 15–30, Jun. 2026, doi: 10.64229/j6d7fr94.

[53] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, pp. 125–146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[54] S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized RWA Infrastructure," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 60–74, Aug. 2026, doi: 10.51903/jtie.v5i2.544.

[55] B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-Style DevOps Copilot for Cloud-Native Troubleshooting: Retrieval-Augmented Runbook Generation and Command-Safety Evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104–118, 2026, doi: 10.51903/jtie.v5i2.534.

[56] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31–47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[57] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74–85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[58] B. Zhou, J. Jin, and D. Zhao, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625–648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[59] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, 2025, doi: 10.51903/jtie.v4i1.485.

[60] D. Zheng, C. Li, and H. Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation," J. Adv. Comput. Syst., vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[61] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75–90, 2026, doi: 10.51903/jtie.v5i2.541.

[62] [62] C. Li, G. Liu, and Z. Zhao, "Cost-Aware LLM-Style Routing for AIOps Log Analysis: Log Parsing, Anomaly Detection, Fault Diagnosis, and Incident Summarization on LogEval Task Files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91–103, 2026, doi: 10.51903/jtie.v5i2.538.

[63] Z. S. Zhong, C. Li, and H. Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting

on ZClawBench," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–360, Apr. 2026, doi: 10.51903/jtie.v5i1.539.

[64] S. He, X. Chang, and E. Sun, "Cross-Cloud Transfer Learning for AI Training Capacity Forecasting under Workload and Topology Distribution Shift," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 100–120, Jan. 2024, doi: 10.69987/JACS.2024.40108.

[65] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 258, pp. 559–567, 2025.

[66] [66] Y. Zhang and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214–229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[67] B. Zhang, Y. Ren, and J. Zou, "LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[68] K. Zhang, Y. Chen, and A. Qian, "Evidence-Grounded Accounting Disclosure Review Cards: A Visual Communication Framework for LLM-Style Explanations over SEC Financial Statements and Notes," Int. J. Graph. Des., vol. 3, no. 2, Oct. 2025, doi: 10.51903/ijgd.v3i2.3710.

[69] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179–185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[70] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216–240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[71] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, pp. 247–264, May 2026, doi: 10.28989/avitec.v8i2.3974.

[72] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[73] Q. Xin, "Host-Based Intrusion Detection with System Call Sequences: Window Localization and Forensic Narratives," AVITEC, vol. 8, no. 2, pp. 325–334, Jun. 2026, doi: 10.28989/avitec.v8i2.3973..