PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 122-131
E-ISSN: 3109-8886

Risk-Controlled Adaptive Multi-Hop RAG with Evidence-Chain Retrieval, Word-Level Hallucination Detection, Attribution, and Calibrated Abstention

[1]Megan Walters, [2]Qing Song, [3]Austin Jiang

[1]Department of Computer Science, University of California, Santa Cruz, Santa Cruz, CA, USA
[2]Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[3]Department of Computer Science, University of California, Riverside, Riverside, CA, USA

Email: [1*]m_walters95@yahoo.com

Abstract

Reliable RAG fails when retrieval omits evidence, generation is unsupported, or confidence is misplaced. We evaluated all 2,556 MultiHop-RAG questions and 2,700 official RAGTruth test responses (2,675 good-quality in primary analysis). On chain-group-disjoint MultiHop-RAG, adaptive retrieval achieved 0.8810 Recall@k and 0.6925 complete-chain rate with 10.85 documents per answerable query; the nearest lower-cost comparator reached 0.8463 and 0.6273. Answer token F1 was 0.5562 and strict URL attribution 0.7753. On RAGTruth, the calibrated detector reached 0.3254 word F1, 0.0986 overlap span F1, 0.00582 ECE, and 0.03550 Brier score. Selective release cut risk from 0.3525 at full coverage to 0.2140 at 80%. Evidence, grounding, provenance, and release risk require separate coordinated measurement.

Keywords: Adaptive Retrieval-Augmented Generation, Multi-Hop Retrieval, Evidence Chains, Hallucination Detection, Source Attribution, Calibration, Selective Prediction, Abstention

Abstrak

Sistem RAG yang andal akan gagal jika tahap pengambilan informasi (retrieval) melewatkan bukti, hasil generasi tidak didukung oleh bukti, atau tingkat kepercayaan yang diberikan tidak tepat. Kami mengevaluasi seluruh 2.556 pertanyaan MultiHop-RAG dan 2.700 respons uji resmi RAGTruth (2.675 di antaranya berkualitas baik untuk analisis utama). Pada dataset MultiHop-RAG dengan rantai bukti yang saling terpisah (chain-group-disjoint), metode pengambilan adaptif mencapai nilai Recall@k sebesar 0,8810 dan tingkat keberhasilan rantai lengkap sebesar 0,6925 dengan rata-rata 10,85 dokumen per kueri yang dapat dijawab; metode pembanding dengan biaya lebih rendah yang terdekat hanya mencapai angka 0,8463 dan 0,6273. Skor F1 untuk token jawaban adalah 0,5562 dan skor atribusi URL ketat adalah 0,7753. Pada dataset RAGTruth, detektor yang telah dikalibrasi mencapai skor F1 kata sebesar 0,3254, F1 rentang tumpang tindih (overlap span) sebesar 0,0986, ECE sebesar 0,00582, dan skor Brier sebesar 0,03550. Strategi pelepasan selektif (selective release) menurunkan risiko dari 0,3525 pada cakupan penuh menjadi 0,2140 pada tingkat cakupan 80%. Aspek bukti, landasan informasi (grounding), asal-usul sumber (provenance), dan risiko pelepasan informasi memerlukan pengukuran terpisah namun terkoordinasi.

***Kata Kunci**: Adaptive Retrieval-Augmented Generation, Pengambilan Informasi Multi-Langkah (Multi-Hop Retrieval), Rantai Bukti, Deteksi Halusinasi, Atribusi Sumber, Kalibrasi, Prediksi Selektif, Abstensi*

A. INTRODUCTION

RAG joins generation and evidence [1], yet retrieved text may not support the answer. MultiHop-RAG distributes evidence across articles [2]; RAGTruth labels unsupported spans [3]. Evidence-chain methods add word detection and provenance [4], while evidence-calibrated QA links coverage to abstention [5]. Reliable RAG must control depth, verify content, attribute sources, and refuse risk.

HotpotQA supervises supporting facts [6], MuSiQue filters shortcuts [7], and Adaptive-RAG conditions retrieval on complexity [8]. IRCoT interleaves retrieval and reasoning [9]; FLARE [10], DRAGIN [11], and Self-RAG [12] retrieve or critique dynamically. Budgeted MultiHop-RAG treats depth as policy [13]. Time-aware memory [14], private preference routing [15], and retrieval-quality prediction [16] extend control to state, privacy, and uncertainty.

BM25 [17], DPR [18], ColBERT [19], and Contriever [20] span sparse, dense, and late-interaction retrieval. We used CPU-feasible BM25 and word/bigram TF-IDF, fused by RRF [21], with pseudo-relevance feedback. Hybrid reranking also appears in financial search [22], code intelligence [23], and conversational recommendation [24]. This isolates fusion and depth without an external model.

RAGAs [25] and ARES [26] score RAG dimensions; FActScore checks atomic claims [27]. Citation completeness [28], attributable-source evaluation [29], SelfCheckGPT [30], HaluEval [31], and KILT [32] address

Diterima Redaksi: 01-08-2026 | Selesai Revisi: xx-xx-2022 | Diterbitkan Online: 07-08-2026

traceability or hallucination. Evidence-chain work adds deterministic provenance [33]. DevOps RAG [34], incident triage [35], and HDFS narratives [36] expose the same operational grounding boundary.

Release is the final boundary. SQuAD 2.0 introduced unanswerable detection [37]; temperature scaling [38] and Brier score [39] assess confidence. Selective prediction reports risk over coverage [40], while conformal risk control selects thresholds under assumptions [41]. Self-verification [42] and calibrated credit rejection [43] show why deferral must preserve safety and utility.

Evidence-grounded systems support DevOps [44], accounting retrieval [45], financial monitoring [46], legal governance [47], humanitarian access [48], and scientific verification [49]. Numerical guardrails separate source consistency from arithmetic correctness [50]. Each requires sufficient retrieval, traceable support, and explicit release; we evaluate depth, grounding, attribution, calibration, and abstention separately.

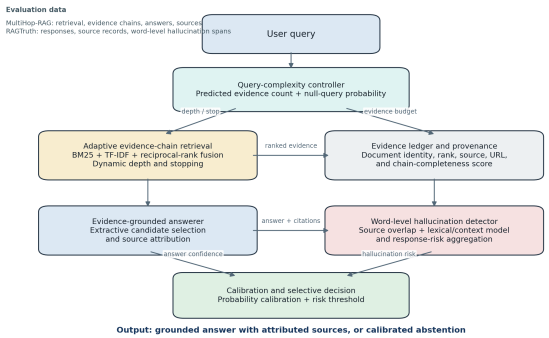

Figure 1 Risk-Controlled Adaptive RAG Architecture And Component-Wise Evaluation Flow.

B. METODE

Data And Label Policy

MultiHop-RAG has 2,556 questions and 609 news articles. All 6,084 evidence records joined the corpus, yielding 5,908 unique question–URL pairs. It contains 856 comparison, 816 inference, 583 temporal, and 301 null questions; answerable items need two to four facts. Gold labels served only supervision or evaluation.

RAGTruth has 2,965 sources and 17,790 responses from six models across QA, data-to-text, and summarization. Its 15,090/2,700 train/test split yielded 14,942/2,675 good responses for primary analysis; every test prediction was retained. Tokens were positive when intervals overlapped a span. Primary labels retained `implicit_true`; sensitivity analysis removed it.

MultiHop-RAG evaluated retrieval, answers, and URL/publisher provenance; RAGTruth evaluated grounding, calibration, and release. Metrics stayed benchmark-specific and met only at the release decision.

Table 1 Datasets, Empirical Units, And Evaluation Roles

Dataset	Primary records	Evidence supervision	Role in the evaluation
MultiHop-RAG	2,556 questions; 609 documents	6,084 fact annotations; 5,908 query–URL pairs	Retrieval, evidence chains, answers, URL and publisher provenance
RAGTruth	17,790 responses; 2,965 sources	14,289 character spans	Word/span detection, calibration, response risk, abstention

Every MultiHop-RAG evidence URL joined to the 609-document corpus; every RAGTruth response joined to one source record.

Table 2 Multihop-RAG Question And Annotation Composition

Dimension	Category	Count	Share
Question type	Comparison	856	33.49%
Question type	Inference	816	31.92%
Question type	Temporal	583	22.81%
Question type	Null	301	11.78%
Annotated fact count	2	1,079	42.21%
Annotated fact count	3	778	30.44%
Annotated fact count	4	398	15.57%
Annotated fact count	0	301	11.78%

Table 3 Ragtruth Official Partitions And Quality Policy

Item	Official train	Official test	Total	Analysis policy
Responses	15,090	2,700	17,790	All test predictions retained
Good responses	14,942	2,675	17,617	Primary metrics
Flagged quality	148	25	173	Incorrect refusal or truncation
Source IDs	2,515	450	2,965	No source crosses official split
Primary word tokens	1,981,518	342,142	2,323,660	Train: 1,586,254 fit + 395,264 validation
Annotated spans	—	—	14,289	Includes 1,928 implicit-true spans

Experimental Splits And Controls

MultiHop-RAG questions were grouped by gold URLs and split with seed 202402 into 1,826/365/365 train/validation/test partitions without query, group, or exact-chain overlap. A query-stratified split tested sensitivity. RAGTruth kept its official boundary; 20% of training source IDs formed source-disjoint validation, keeping six responses together. Test labels were used once.

Runs used fixed seeds, deterministic ties, per-instance predictions, and timing. Confidence intervals used 500 evidence-chain-group or response bootstrap resamples; Table IV gives fixed settings.

Table 4 Fixed Experimental Configuration

Component	Setting	Selection or fitting control
MultiHop split	1,826 / 365 / 365	Stratified evidence-chain groups; seed 202402
BM25	k1 = 1.5; b = 0.75	All 609 documents
TF-IDF / RRF	Word 1–2 grams; RRF c = 60	Fixed depths {2, 4, 6, 8, 10, 12}
Feedback	Top 2 documents; weight 0.35	RRF of BM25, TF-IDF, and feedback rank
Complexity	Balanced multinomial logistic regression; C = 2	Question text only; null threshold on validation
Sufficiency	250-tree random forest; depth 12; leaf 5	Threshold 0.90 selected on validation
Answerer	TF-IDF nearest training question + text constraint	Retrieved capitalized-span fallback
Word detector	180,000 TF-IDF features; SGD log loss	alpha = 10^-6; 20 iterations; balanced classes
Calibration	Platt, isotonic, Evidence-Stacked; ECE-15	Source-disjoint RAGTruth validation
Uncertainty	500 bootstrap samples	Evidence-chain-group/response resampling; 95% intervals
Risk control	Targets 0.10, 0.20, 0.30	One-sided 95% Clopper–Pearson validation bound

Retrieval And Dynamic Evidence Depth

Documents combined metadata and body. BM25 and word/bigram TF-IDF ranked all 609 documents; RRF used constant 60. Feedback averaged the top two RRF document vectors, weighted this view 0.35, reranked, and fused all three rankings. Fixed depths were 2, 4, 6, 8, 10, and 12.

A query-only classifier predicted budget class 0, 2, 3, or 4. Null probability controlled zero-document return; otherwise the class set initial depth. A second model estimated complete-chain sufficiency from training retrieval states. The controller added documents until the validation threshold or depth 12, using observed state rather than gold count.

Retrieval metrics were evidence Recall@k, binary nDCG, precision, any-evidence success, complete-chain rate, and documents read. Null metrics included F1, AUROC, average precision, ECE, and Brier score. Validation chose the largest fixed depth not exceeding adaptive mean cost; depth 12 was the high-cost ceiling.

Evidence-grounded answering and provenance

The answerer ranked training answers by query similarity, accepting non-binary candidates only when found in retrieved evidence. Yes/no questions used nearest labels; others used a deterministic capitalized-span extractor. Null decisions emitted “Insufficient information”; held-out answers never entered prediction.

EM and token F1 used SQuAD normalization. Gold-string coverage was an extractive ceiling. Each non-null answer named one article: strict attribution required a gold URL; source attribution compared publishers. Correct null abstentions received correct no-citation provenance.

Word-level hallucination detection

RAGTruth offsets aligned tokens to spans. A lexical baseline flagged source-absent content and low-overlap neighborhoods. The learned class-weighted linear detector used token/context n-grams, shape, source frequency, local support, position, task, generator, collection, and temperature. Source-disjoint validation selected calibration and threshold.

Word metrics were micro-averaged. Exact span F1 required identical boundaries; overlap F1 used one-to-one token IoU ≥0.5. Response risk aggregated top token probabilities so localized spans mattered without length dominating. We also reported discrimination, subgroup scores, and sensitivity without `implicit_true`; training never crossed validation.

Calibration, attribution proxy, and selective release

Validation fitted Platt, isotonic, and Evidence-Stacked calibration, the last combining detector logit and support. This follows late fusion [51] and parallels PD calibration [52] and cluster uncertainty control [53], though here probability only governed release. ECE and Brier assessed test calibration. Validation chose Clopper–Pearson-bounded target-risk thresholds, then fixed them for test. Reusing validation labels makes bounds selection diagnostics, not post-selection guarantees.

RAGTruth lacks gold evidence-unit IDs. We matched response sentences to passage, sentence, or record units as a grounding proxy, not attribution accuracy. Strict provenance remained confined to MultiHop-RAG URLs and publishers; evidence-card interfaces likewise must not turn a plausible match into benchmark gold [54].

C. RESULTS AND DISCUSSION

Retrieval effectiveness and evidence-chain completion

Fusion consistently exceeded BM25. At depth 8, RRF reached 0.8396 recall and 0.7613 nDCG, while TF-IDF had slightly higher complete-chain rate (0.6262 versus 0.6160). At depth 12, RRF reached 0.8984 recall and 0.7375 complete chains, confirming that average rank quality and full-chain recovery differ.

On chain-group-disjoint test data, adaptive PRF-RRF achieved 0.8810 recall, 0.7716 nDCG, and 0.6925 complete chains. Cluster-bootstrap gains over PRF-RRF@10 were 0.0347, 0.0126, and 0.0652; differences from @12 were small and their intervals reached zero. The controller occupied a useful point between fixed budgets.

Table 5 Fixed-Depth Retrieval On All Answerable Multihop-RAG Questions

System	Depth	Questions	Recall	nDCG	Complete chain	Evidence precision
BM25	8	2,255	0.8155	0.7331	0.5814	0.2600
TF-IDF	8	2,255	0.8384	0.7606	0.6262	0.2676
RRF	8	2,255	0.8396	0.7613	0.6160	0.2681
BM25	12	2,255	0.8773	0.7577	0.6967	0.1884
TF-IDF	12	2,255	0.8940	0.7827	0.7220	0.1920
RRF	12	2,255	0.8984	0.7847	0.7375	0.1931

Fixed rankings require no learned query controller; metrics deduplicate multiple annotated facts from the same URL.

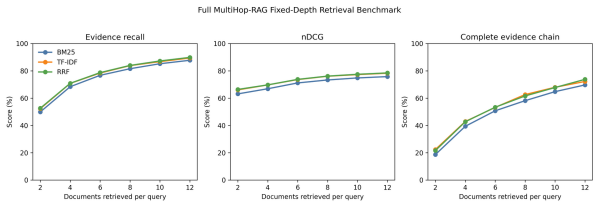

Figure 2 Fixed Retrieval Performance Across Depths 2, 4, 6, 8, 10, And 12 On All Answerable Multihop-RAG Questions.

Table 6 RRF At Depth 8 By Multihop-RAG Question Type

Question type	Answerable n	Recall	nDCG	Complete chain	Any evidence
Comparison	856	0.8865	0.8008	0.7640	0.9965
Inference	816	0.7734	0.7158	0.4167	0.9975
Temporal	583	0.8634	0.7672	0.6775	0.9897

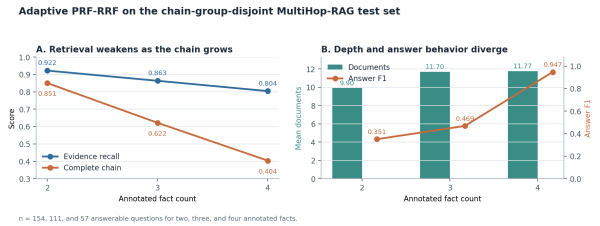

Figure 3 Adaptive retrieval and answering by annotated fact count on the chain-group-disjoint test set.

Dynamic Depth, Null Detection, And Resource Use

The controller achieved 0.7589 budget accuracy and 0.9885 null F1. At sufficiency 0.90 it read 10.854 documents per answerable query, 0.854 more than depth 10 and 1.146 fewer than 12. Hybrid-cloud serving ties context to latency/cost [55]; conformal GPU envelopes likewise connect uncertain resources to bounded decisions [56].

The query-stratified sensitivity split produced 0.8766 recall, 0.7602 nDCG, 0.6844 complete chains, and 10.608 documents, close to primary results. Near-perfect null discrimination remains benchmark-specific because insufficient-information items contain regular textual cues.

Table 7 Adaptive Retrieval, Fixed Comparators, And Split Sensitivity

Split	System	Docs / answerable	Recall	nDCG	Complete chain	Null F1
Chain-disjoint	PRF-RRF@10	10.000	0.8463	0.7590	0.6273	—
Chain-disjoint	Adaptive PRF-RRF	10.854	0.8810	0.7716	0.6925	0.9885
Chain-disjoint	PRF-RRF@12	12.000	0.8864	0.7743	0.6988	—
Query-stratified	Adaptive PRF-RRF	10.608	0.8766	0.7602	0.6844	1.0000

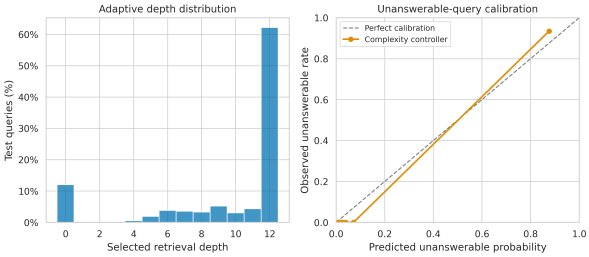

Figure 4 Adaptive Depth Distribution And Null-Controller Calibration On The Chain-Group-Disjoint Split.

Answer Correctness And Source Attribution

Across 365 test questions, adaptive answer EM/F1 was 0.5562, including 43 correct null abstentions; the 322 answerable questions scored 0.4969. Gold strings occurred in 0.9697 of 132 eligible retrieved evidence sets, so selection and composition—not retrieval alone—were the main answer bottleneck.

Adaptive URL attribution was 0.7753 versus 0.9397 for publisher identity; answerable-only values were 0.7453 and 0.9317. Related articles from the correct publisher often failed the exact required hop. Scientific-search [54] and accounting-review evidence cards [57] likewise motivate displaying the exact support unit rather than only a plausible source family.

Table 8 Held-Out Answer And Provenance Results On The Chain-Disjoint Test Set

System	Answer string cov.	EM	Token F1	URL accuracy	Publisher accuracy	Source-chain complete
BM25@8	0.9773	0.5479	0.5498	0.7260	0.8822	0.8944
TF-IDF@8	0.9697	0.5589	0.5589	0.7836	0.9534	0.9193
RRF@8	0.9773	0.5534	0.5534	0.7808	0.9315	0.9068
PRF-RRF@8	0.9697	0.5562	0.5562	0.7753	0.9370	0.9161
Adaptive PRF-RRF	0.9697	0.5562	0.5562	0.7753	0.9397	0.9565

Answer and attribution columns include 43 null queries; answerable-only adaptive EM/F1 was 0.4969. Answer-string coverage uses 132 eligible non-binary answerable questions.

Word-level and span-level hallucination detection

Among 342,142 eligible RAGTruth tokens, 4.20% were positive. Word F1 rose from 0.1305 for the lexical baseline to 0.3198 for TF-IDF and 0.3254 for Evidence-Stacked (0.2702 precision, 0.4091 recall; bootstrap interval 0.3045–0.3474). Excluding `implicit_true` yielded 0.3030, so that label did not drive the result.

Localization was harder: exact span F1 was 0.0305 and overlap F1 0.0986. Word F1 was 0.3699 for data-to-text, 0.3316 for QA, and 0.1829 for summarization; generator F1 ranged 0.0277–0.3673 as prevalence changed. This is analogous—not equivalent—to localizing anomalous windows before forensic narratives [58]; incident cards likewise keep evidence inspectable [59].

Table 9 Word-Level Context-Grounding Detection On Ragtruth

Detector	Precision	Recall	F1	AUROC	AUPRC	Brier	ECE-15	Span F1 (IoU ≥0.5)
Lexical-Evidence	0.0867	0.2639	0.1305	0.6635	0.0712	0.19088	0.21490	0.0150
TFIDF-Linear	0.2653	0.4027	0.3198	0.8350	0.2446	0.13645	0.23104	0.0954
TFIDF-Platt	0.2653	0.4027	0.3198	0.8350	0.2446	0.03559	0.00601	0.0954
TFIDF-Isotonic	0.2653	0.4027	0.3198	0.8349	0.2387	0.03560	0.00606	0.0954
Evidence-Stacked	0.2702	0.4091	0.3254	0.8359	0.2468	0.03550	0.00582	0.0986

Primary rule includes implicit-true labels; Evidence-Stacked exact-boundary span F1 was 0.0305. Word-F1 95% bootstrap CI: [0.3045, 0.3474].

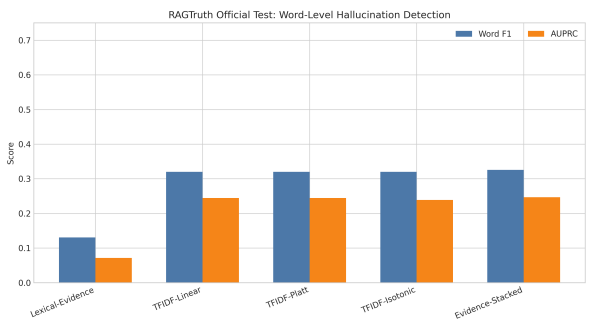

Figure 5 Ragtruth Word-Level Detector Comparison Of F1 And Area Under The Precision--Recall Curve.

Table 10 Evidence-Stacked Word Detection By Task And Generator

Dimension	Group	Tokens	Positive rate	Precision	Recall	F1
Task	Data2txt	143,038	0.0427	0.3162	0.4457	0.3699
Task	QA	99,665	0.0533	0.2471	0.5037	0.3316
Task	Summary	99,439	0.0298	0.2075	0.1635	0.1829
Generator	GPT-3.5-Turbo	63,236	0.0104	0.0791	0.0168	0.0277
Generator	GPT-4	50,419	0.0061	0.1619	0.0552	0.0823
Generator	Llama-2-13B	61,118	0.0537	0.2828	0.4855	0.3574
Generator	Llama-2-70B	57,356	0.0497	0.2533	0.3517	0.2945
Generator	Llama-2-7B	59,084	0.0620	0.2469	0.4279	0.3131
Generator	Mistral-7B	50,929	0.0712	0.3026	0.4670	0.3673

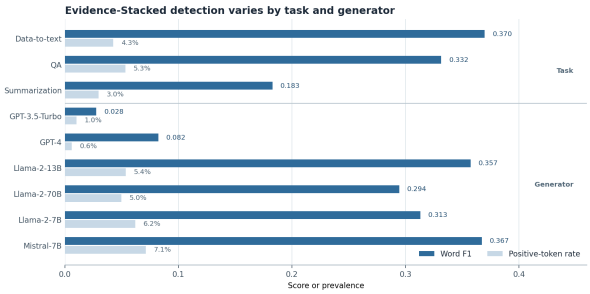

Figure 6 Evidence-Stacked Word F1 By Task And Generator, With Positive-Token Prevalence Shown For Context.

Calibration And Selective Abstention

Platt scaling reduced word-level Brier score from 0.13645 to 0.03559 and ECE from 0.23104 to 0.00601; Evidence-Stacked reached 0.03550 and 0.00582. At response level it achieved 0.7425 F1, 0.8940 AUROC, 0.1292 Brier, and 0.0738 ECE, so operating points were selected on held-out sources.

Risk fell from 0.3525 at full coverage to 0.2140/0.1624/0.1202/0.0823 at 80/70/60/50%. Validation targets 0.10/0.20/0.30 gave test risks 0.0369/0.1095/0.2010. This parallels calibrated credit deferral [60] and cost-aware AIOps routing [61]. Bounds are empirical diagnostics, not post-selection guarantees.

Table 11 Token- And Response-Level Discrimination And Calibration

Level	Model	F1	AUROC	AUPRC	Brier	ECE-15	Threshold
Word	TFIDF-Linear	0.3198	0.8350	0.2446	0.13645	0.23104	0.81094
Word	TFIDF-Platt	0.3198	0.8350	0.2446	0.03559	0.00601	0.17100
Word	TFIDF-Isotonic	0.3198	0.8349	0.2387	0.03560	0.00606	0.19176

Level	Model	F1	AUROC	AUPRC	Brier	ECE-15	Threshold
Word	Evidence-Stacked	0.3254	0.8359	0.2468	0.03550	0.00582	0.17295
Response	Lexical-Evidence	0.7038	0.8693	0.7925	0.14464	0.07980	0.41311
Response	TFIDF-Platt	0.7356	0.8921	0.8366	0.13033	0.07344	0.38442
Response	Evidence-Stacked	0.7425	0.8940	0.8402	0.12919	0.07382	0.40056

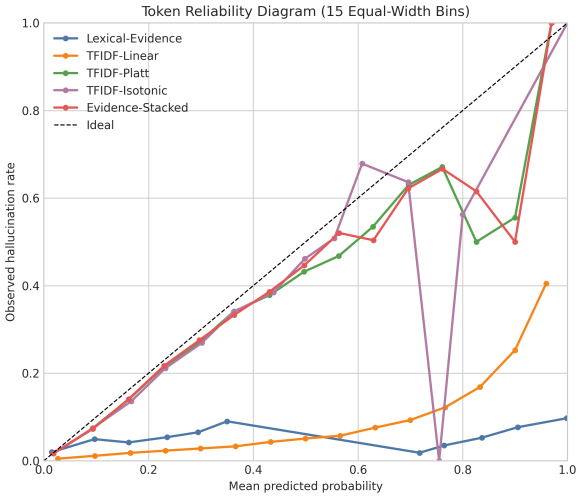

Figure 7 Word-Level Calibration Curves For The Ragtruth Detectors; The Diagonal Denotes Ideal Calibration.

Table 12 Evidence-Stacked Response Risk Under Selective Release

Rule	Threshold	Coverage	Accepted n	Test risk	Validation upper bound
Full release	0.9910	1.0000	2,675	0.3525	—
Coverage slice	0.7818	0.8000	2,140	0.2140	—
Coverage slice	0.6157	0.6998	1,872	0.1624	—
Coverage slice	0.4586	0.6000	1,605	0.1202	—
Coverage slice	0.3550	0.4998	1,337	0.0823	—
Target risk 0.10	0.2016	0.3039	813	0.0369	0.0992
Target risk 0.20	0.4333	0.5802	1,552	0.1095	0.1999
Target risk 0.30	0.7457	0.7738	2,070	0.2010	0.2997

Target-risk thresholds are validation-selected operating points; the displayed unadjusted one-sided bounds are not post-selection guarantees.

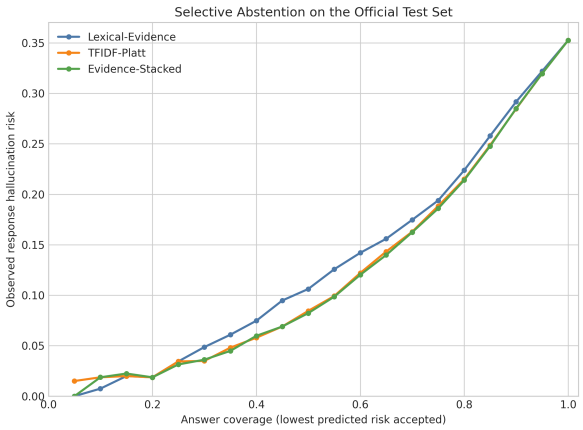

Figure 8 Response-Level Hallucination Risk As A Function Of Released Coverage For Three Scoring Models.

Error analysis and implications

Retrieval errors concentrated in weak lexical bridges. RRF@8 chain completion was 0.7640 for comparison, 0.6775 for temporal, and 0.4167 for inference questions. Adaptive inference recall/chain rate was 0.8319/0.5299 versus 0.9126/0.8033 for comparison; chain rate fell from 0.8506 for two facts to 0.4035 for four.

Answer errors differed: temporal questions had 0.9036 retrieval recall but 0.0964 answer F1, whereas inference had 0.8319 and 0.8803. The lexical answerer did not reliably compose temporal relations, confirming that retrieval and answer quality need separate scores.

RAGTruth errors mixed low-prevalence groups, long-clause boundaries, and ambiguous lexical matches. The source-unit score was a grounding proxy (AUROC 0.6834, AUPRC 0.1263), not attribution accuracy. Retrieved OpenStack prototypes [62], incident cards [59], AIOps routing [61], and trajectory reliability prediction [63] preserve inspectable local signals; none substitutes for gold passage links.

Reliability spans evidence recall, chain sufficiency, traceable use, calibrated risk, and inspectable release. Prompt-injection cards [64] and refusal [65] expose escalation; review-grounded recommendation [66], counseling support [67], and infrastructure risk briefs [68] expose evidence or uncertainty. Interfaces do not replace attribution.

Scope is bounded: 609 English news articles, an extractive answerer, six RAGTruth model families, and no gold citation links. Calibration used the official distribution. Red-teaming [69] and conformal incident alerts [70] show thresholds need refresh after shift. Results are within-dataset, not universal across domains, languages, or generators.

D. CONCLUSIONS

On MultiHop-RAG, sparse fusion and feedback strengthened fixed retrieval; the controller improved complete-chain recovery over depth 10 and approached depth 12 with fewer documents. Strict URL attribution

exposed errors hidden by answer-only metrics. On RAGTruth, lexical-context detection improved word/span scores over source overlap, and calibration made imperfect scores useful for selective release.

Reliable RAG should record four objects per answer: evidence chain, answer text, source set, and calibrated release probability. Evaluation should preserve them through ranking, chain completeness, answer agreement, attribution, span localization, and risk–coverage. Answer accuracy alone cannot distinguish missing evidence, unsupported generation, or overconfident release.

E. REFERENCES

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

[2] Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries," in Proc. 1st Conf. Language Modeling (COLM), 2024, doi: 10.48550/arXiv.2401.15391.

[3] C. Niu et al., "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics, 2024, pp. 10862–10878, doi: 10.18653/v1/2024.acl-long.585.

[4] 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.

[5] [5] 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.

[6] [6] Z. Yang et al., "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," in Proc. 2018 Conf. Empirical Methods Natural Language Processing, 2018, pp. 2369–2380, doi: 10.18653/v1/D18-1259.

[7] [7] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "MuSiQue: Multihop Questions via Single-hop Question Composition," Trans. Assoc. Comput. Linguistics, vol. 10, pp. 539–554, 2022, doi: 10.1162/tacl_a_00475.

[8] [8] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," in Proc. NAACL, 2024, pp. 7036–7050, doi: 10.18653/v1/2024.naacl-long.389.

[9] [9] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions," in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics, 2023, pp. 10014–10037, doi: 10.18653/v1/2023.acl-long.557.

[10] [10] Z. Jiang et al., "Active Retrieval Augmented Generation," in Proc. 2023 Conf. Empirical Methods Natural Language Processing, 2023, pp. 7969–7992, doi: 10.18653/v1/2023.emnlp-main.495.

[11] [11] W. Su, Y. Tang, Q. Ai, Z. Wu, and Y. Liu, "DRAGIN: Dynamic Retrieval Augmented Generation Based on the Real-Time Information Needs of Large Language Models," in Proc. ACL, 2024, pp. 12991–13013, doi: 10.18653/v1/2024.acl-long.702.

[12] [12] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in Proc. ICLR, 2024.

[13] [13] 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.

[14] [14] 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.

[15] [15] 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.

[16] [16] 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.

[17] [17] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[18] [18] 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.

[19] [19] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proc. 43rd Int. ACM SIGIR Conf., 2020, pp. 39–48, doi: 10.1145/3397271.3401075.

[20] [20] G. Izacard et al., "Unsupervised Dense Information Retrieval with Contrastive Learning," Trans. Mach. Learn. Res., 2022.

[21] [21] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,"

in Proc. 32nd Int. ACM SIGIR Conf., 2009, pp. 758–759, doi: 10.1145/1571941.1572114.

[22] [22] 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.

[23] [23] 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.

[24] [24] Q. Xin, “Behavior Retrieval plus Response Generation for Interpretable Conversational Personalized Recommendation,” IJEEPSE, vol. 9, no. 2, pp. 120–136, Jul. 2026, doi: 10.31258/ijeepse.9.2.120-136.

[25] [25] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAs: Automated Evaluation of Retrieval Augmented Generation,” in Proc. EACL System Demonstrations, 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16.

[26] [26] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, “ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems,” in Proc. NAACL, 2024, pp. 338–354, doi: 10.18653/v1/2024.naacl-long.20.

[27] [27] S. Min et al., “FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation,” in Proc. EMNLP, 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741.

[28] [28] T. Gao, H. Yen, J. Yu, and D. Chen, “Enabling Large Language Models to Generate Text with Citations,” in Proc. EMNLP, 2023, pp. 6465–6488, doi: 10.18653/v1/2023.emnlp-main.398.

[29] [29] H. Rashkin et al., “Measuring Attribution in Natural Language Generation Models,” Comput. Linguistics, vol. 49, no. 4, pp. 777–840, 2023, doi: 10.1162/coli_a_00486.

[30] [30] P. Manakul, A. Liusie, and M. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” in Proc. EMNLP, 2023, pp. 9004–9017, doi: 10.18653/v1/2023.emnlp-main.557.

[31] [31] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,” in Proc. EMNLP, 2023, pp. 6449–6464, doi: 10.18653/v1/2023.emnlp-main.397.

[32] [32] F. Petroni et al., “KILT: A Benchmark for Knowledge Intensive Language Tasks,” in Proc. NAACL, 2021, pp. 2523–2544, doi: 10.18653/v1/2021.naacl-main.200.

[33] [33] 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.

[34] [34] 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.

[35] [35] 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.

[36] [36] 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.

[37] [37] P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD,” in Proc. ACL, 2018, pp. 784–789, doi: 10.18653/v1/P18-2124.

[38] [38] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in Proc. 34th Int. Conf. Machine Learning, vol. 70, 2017, pp. 1321–1330.

[39] [39] G. W. Brier, “Verification of Forecasts Expressed in Terms of Probability,” Mon. Weather Rev., vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[40] [40] Y. Geifman and R. El-Yaniv, “SelectiveNet: A Deep Neural Network with an Integrated Reject Option,” in Proc. 36th Int. Conf. Machine Learning, vol. 97, 2019, pp. 2151–2159.

[41] [41] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, “Conformal Risk Control,” in Proc. ICLR, 2024.

[42] [42] 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.

[43] [43] Y. Chen et al., “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.

[44] [44] 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, Jun. 2026, doi: 10.51903/jtie.v5i2.534.

[45] [45] 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.

[46] [46] 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, Jun. 2026, doi: 10.51903/jtie.v5i2.544.

[47] [47] J. Li and A. Zhou, "Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces," Rule Law Stud. J., vol. 2, no. 2, pp. 91–108, Jun. 2026, doi: 10.64780/rolsj.v2i2.225.

[48] [48] 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.

[49] [49] 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.

[50] [50] 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, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[51] [51] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception: Confidence Calibration and Fusion Rule Learning," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[52] [52] 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.

[53] [53] S. He, H. Tu, and I. Liu, "Safe PD Capacity Forecasting with Time-Series Foundation Models and Calibrated Uncertainty for Heterogeneous GPU Clusters," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 48–66, Apr. 2023, doi: 10.69987/JACS.2023.30404.

[54] [54] 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.

[55] [55] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182–2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.

[56] [56] S. Chen, S. He, and E. Sun, "Risk-Bounded GPU Resource Oversubscription via Conformal Demand Envelopes in Production AI Clusters," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 119–134, May 2024, doi: 10.69987/JACS.2024.40509.

[57] [57] 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, p. 395, Oct. 2025. [Online]. Available: https://journal.stekom.ac.id/index.php/ijgd/article/view/3710

[58] [58] 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.

[59] [59] 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.

[60] [60] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," J-INTECH (J. Inf. Technol.), vol. 14, no. 2, pp. 215–231, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.

[61] [61] 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, Jun. 2026, doi: 10.51903/jtie.v5i2.538.

[62] [62] 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.

[63] [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] [64] 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.

[65] [65] D. Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal and Utility Preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[66] [66] 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, May 2026, doi: 10.54732/jeecs.v11i1.2.

[67] [67] 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.

[68] [68] G. Liu, S. He, and H. Wong, "LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–213, May 2025, doi: 10.51903/ijgd.v3i1.3723.

[69] [69] 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.

[70] [70] 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.