Journal of Knowledge Learning and Science Technology
ISSN: 2959-6386 (Online)
2026, Vol. 5, No. 2, pp. 79–102
DOI: https://doi.org/10.60087/jklst.vol5.n2.006
Journal homepage: https://jklst.org/index.php/home

Budgeted Multi-Hop Retrieval Agents with Evidence-Chain Reliability and Cost-Aware Abstention

Caleb Kong
Computer Science, Rutgers University, New Brunswick, NJ
**Corresponding author: Caleb Kong, ckong_research@outlook.com*

Abstract

Fixed top-k retrieval spends the same context budget on every question without indicating whether a multi-document evidence chain is complete. This study evaluates a budgeted agent on MultiHop-RAG using all 609 documents and 2,556 questions. The 1,534/511/511 train/validation/test split kept each evidence group in one partition. The pipeline compared BM25, 256-dimensional latent-semantic retrieval, equal and weighted reciprocal rank fusion, supervised reranking, fixed depth, and calibrated dynamic stopping. A sparse reader, null detector, and isotonic calibration supported answer, evidence, citation, cost, and abstention evaluation. On 451 answerable test questions, reranking achieved 0.9069 Recall@10, 0.8064 nDCG@10, and 0.7428 complete-chain retrieval. The dynamic agent used a mean of 544.16 tokens and obtained 0.6184 F1, compared with 666.29 tokens and 0.6243 F1 for the null-gated Reranker-8 policy. Its paired F1 difference was -0.0059 with a 95% cluster-bootstrap interval of [-0.0246, 0.0115], while the 122.13-token reduction remained significant after Holm adjustment. Isotonic calibration reduced expected calibration error from 0.2671 to 0.0321. Selective answering operated at 0.3326 coverage and 0.2200 exact-match risk. Snippet retention and citation precision, rather than candidate generation alone, limited joint reliability. The results establish a measured quality-cost trade-off and identify where deeper retrieval still fails to improve answers.

Keywords: *—multi-hop retrieval, retrieval-augmented generation, evidence chains, reciprocal rank fusion, dynamic stopping, calibration, selective prediction, abstention.*

Article information: Received: 02/03/2026; Accepted: 01/06/2026; Published: 25/06/2026

1. Introduction

Retrieval-augmented generation separates an answering model from an external document collection so responses can use current, inspectable evidence rather than only parametric memory [1], [2. Multi-hop settings make this separation harder because a complete answer depends on a connected source set, not merely one relevant passage. Evidence-chain attribution has therefore become a central reliability objective [3], while calibrated refusal on unanswerable questions [4] and multilingual information access [5] show that retrieval adequacy must be estimated before fluent generation is trusted. One missing document can break the chain; one distractor can consume context or redirect the reader.

MultiHop-RAG provides a direct setting for studying these decisions. Its news corpus contains entity, comparison, temporal, and inference questions that require two to four documents,

alongside null cases with no valid chain [1]. A prior budgeted retrieval-policy evaluation on the same benchmark [6] established the relevance of cost-sensitive multi-hop control. The present study extends that line with grouped leakage prevention, local sparse and dense fusion, a supervised reranker, prefix-level chain calibration, fact-retention and citation diagnostics, selective abstention, and paired evidence-group inference.

This study asks three linked questions. First, how do lexical retrieval, a local dense representation, rank fusion, and supervised reranking differ in evidence recall and complete-chain recovery? Second, how much context cost can a calibrated stopping policy remove relative to fixed deep retrieval without a material answer-quality loss, a control problem related to adaptive routing [7] and budgeted model cascades [8]? Third, does a reliability score combining answer, chain-completion, and null probabilities support useful abstention? Tool-use trajectory forecasting [9] motivates the separation between prefix reliability and final task success.

The study provides an end-to-end empirical account using the complete benchmark, an evidence-group split, validation-only parameter selection, and held-out testing. It reports ranking, answer, evidence, citation, cost, calibration, risk-coverage, hop, and paired-bootstrap results. These measurements separate evidence acquisition from answer conversion and expose the inputs needed by transparent scientific search displays [10] and numerical guardrail systems [11]: reranking improves document recovery, while the reader and snippet selector limit the value of additional depth.

2. Literature Review

A. Multi-Hop Benchmarks

HotpotQA established multi-hop question answering with supporting facts and bridge and comparison questions [12]. 2WikiMultiHopQA added explicit paths [13], while MuSiQue reduced disconnected shortcuts through dependency-controlled composition [14]. MultiHop-RAG instead couples news documents and metadata with multi-document answers and null cases [1]. Because benchmark behavior need not transfer unchanged, work on approximately shared cross-domain features [15] supports treating the grouped held-out split as an internal-validity design rather than evidence of universal generalization.

B. Retrieval and Evidence-Chain Construction

Multi-hop dense retrieval is sequential [16]: later searches condition on the question and recovered evidence. IRCoT interleaves retrieval with partial reasoning [17], and EfficientRAG uses smaller query and filtering components [18]. Generative multi-hop retrieval decodes target passages [19], whereas iterative retrieval-generation updates search from current output [20]. Evidence-chain provenance [3] and learned prediction of RAG retrieval quality [21] address the corresponding question of whether the recovered state is dependable. These approaches exploit intermediate state but can propagate an early distractor.

Lexical and semantic retrievers provide complementary signals. BM^{25} remains strong for exact entities and events [22]; dense passage retrieval recovers semantic matches [23]; and latent semantic analysis offers a smaller local projection [24]. BEIR shows that rankings vary across domains [25]. Reciprocal rank fusion combines heterogeneous lists without comparable raw

scores [26], while spectral rank aggregation provides a complementary theoretical view [27]. Two-stage interaction models rerank a candidate pool [28]; hybrid query rewriting and listwise ranking on FiQA [29] further illustrate the layered retrieve-fuse-rerank pattern used here.

C. Adaptive Retrieval and Stopping

Adaptive RAG replaces a fixed schedule with decisions driven by query or generation state. Self-RAG learns retrieval and reflection tokens [30]; FLARE triggers retrieval under low confidence [31]; Adaptive-RAG routes by predicted complexity [7]; and DRAGIN models when and what to retrieve [32]. Cost-aware AIOps routing similarly assigns cases to different analysis paths [33], although its utility target differs from evidence-chain completeness. A stopping controller must therefore estimate both the marginal value of another document and whether accumulated support already satisfies the question.

D. Calibration, Abstention, and Grounded Evaluation

Stopping and refusal require probabilities that correspond to observed correctness. Modern predictive models can be strongly miscalibrated [34]; selective-classification theory formalizes abstention through coverage and risk [35], with confidence-threshold implementations [36] and selective QA under shift [37]. Applied systems calibrate different objects: cross-sensor fusion confidence [38], resume-pair screening and refusal [39], spectral-estimator uncertainty [40], and counterfactual credit-risk decisions [41]. For multi-hop retrieval, the reliability variable must represent evidence sufficiency as well as answer plausibility.

Evaluation must separate ranking, answering, grounding, and cost. nDCG rewards early evidence [42]; normalized exact match and token F^1 capture answer strings [43]; and citation-centered evaluation tests support and attribution [44]. RAGAS [45] and ARES [46] decompose context and answer quality. FrugalGPT adds budgeted routing [8]. Applied to multi-hop retrieval, the analogous decision is whether another hop justifies its context while improving complete support rather than merely adding a plausible document.

E. Evidence-Grounded Systems and Operational Transfer

Evidence-constrained retrieval has moved into domains where an unsupported answer carries operational consequences. Accounting-aware due-diligence retrieval [47], disclosure review cards [48], and RWA monitoring agents [49] tie outputs to source records. Climate claim verification [50] and multi-regulation product counsel [51] likewise require ranked evidence and an explicit path to non-support. These studies motivate treating chain coverage, citation accuracy, and abstention as operational endpoints rather than stylistic qualities of generated prose.

The same distinction shapes human-facing evidence displays. Financial risk dashboards [52], medical explanation cards [53], patient safety communication [54], medical-response interface benchmarks [55], and e-commerce recommendation cards [56] all depend on visible evidence and uncertainty. Work aligning LLM design criticism with human judgment [57] further shows that explanation quality must be inspectable. The present study does not run a UI experiment; it measures retrieval, support, and confidence prerequisites that such interfaces would consume.

Evidence-grounded assistance also spans recommendation, counseling, and education. Review-grounded recommendation with faithfulness testing [58] and a therapist-facing retrieval copilot

[59] connect ranked context to a controlled output. Conversational recommendation retrieves prior behavior [60], counseling interfaces link support cards to evidence [61], and rubric-grounded essay feedback makes scoring support auditable [62]. Interface-microcopy audits explain dark patterns [63], while student-retention rationales [64] separate a prediction from the evidence shown to a human decision maker.

Operational log systems expose a closely related conversion problem: detecting an event is not equivalent to grounding an actionable explanation. HDFS failure narratives [65], retrieved OpenStack prototypes [66], cost-aware AIOps routing [33], and incident visualization cards [67] each connect selection to explanation. Conformal log alert control [68], interpretable CloudTrail stage detection [69], and self-supervised LogBERT-style detection [70] add uncertainty, sequence structure, or representation learning. Together they reinforce the present separation of candidate recovery, retained fact support, answer conversion, and escalation.

Security studies provide a second reason not to equate model confidence with trustworthy evidence. Language-guided DDoS detection [71], IoT anomaly representation [72], and predictive attack mitigation [73] focus on identifying abnormal behavior; system-call localization with forensic narratives [74] adds an evidence window. Risk cards for prompt-injected agents [75] make the oversight requirement explicit. Although the benchmark studied here is not adversarial, its strict citation metric similarly prevents a high-confidence answer from masking unsupported selected context.

Context cost also has a deployment meaning beyond token count. Multi-horizon GPU forecasting [76], cold-start tenant prediction [77], and hybrid-cloud LLM architecture [78] characterize uncertain demand and placement. Profit-aware GPU admission [79], power-aware inventory planning [80], capacity risk cards [81], and GPU-memory prediction [82] turn forecasts into resource decisions. These systems are not experimental baselines for this paper; they motivate reporting an explicit quality-cost frontier instead of optimizing answer quality alone.

Finally, dynamic stopping is a learned policy and therefore benefits from the discipline used in offline decision studies. Counterfactual slot evaluation [83], privacy-robust incrementality [84], conservative off-policy selection [85], constrained marketing allocation [86], and identifier-safe budget optimization [87] all distinguish model fit from policy value. Text-grounded advertising rationales [88], visual brief cards [89], and designer-editable decision cards [90] illustrate how policy inputs and outputs can remain inspectable. Here, held-out policy comparisons and paired bootstrap intervals serve that evaluative role.

3. Methodology

A. Benchmark Audit and Partitioning

The experiments used the MultiHop-RAG benchmark [1]. The corpus file contained 609 unique news documents and occupied 6,785,567 bytes; the question file occupied 5,171,312 bytes and contained 2,556 questions. There were 2,255 answerable questions and 301 null questions, with 6,084 evidence annotations. Every evidence URL and title matched one corpus document, and every annotated fact was an exact substring of the referenced body. The answerable subset contained 1,169 two-document, 774 three-document, and 312 four-document questions.

Document-level retrieval metrics deduplicated repeated URLs within a question, while fact coverage retained all annotations. Table I summarizes the dataset and integrity checks.

TABLE I. DATASET AUDIT AND INTEGRITY

Statistic	Measured value
Corpus documents	609
Questions	2,556
Answerable / null	2,255 / 301
Evidence annotations	6,084
Two / three / four gold documents	1,169 / 774 / 312
Comparison / inference / temporal / null	856 / 816 / 583 / 301
Evidence URL and title match	100.00%
Evidence fact substring match	100.00%
Corpus / question bytes	6,785,567 / 5,171,312

Questions sharing the same sorted set of evidence URLs were assigned to one canonical group; each null question received its own group. Five shuffled stratified group folds used seed 42 and preserved question type together with gold-document count. Folds 0–2 formed training, fold 3 validation, and fold 4 test, producing 1,534/511/511 questions. No evidence group crossed partitions. Question type was used only for stratification and reporting, never as a model feature. Fusion weights, thresholds, and calibrators were selected on validation data, and final comparisons used the test split. Table II gives the question-type distribution in each partition.

TABLE II. GROUPED SPLIT DISTRIBUTION

Split	Comparison	Inference	Temporal	Null	Total
Training	512	490	351	181	1,534
Validation	172	163	116	60	511
Test	172	163	116	60	511

B. Retrieval, Fusion, and Reranking

Each indexed document concatenated title twice, source, category, publication date, and body. Title repetition increased headline influence without changing source content. A case-insensitive regular expression extracted alphanumeric tokens with internal apostrophes. BM25 used a 50,000-feature sparse term-frequency matrix, k1 = 1.2, b = 0.75, and binary query terms [22].

$$BM25(q,d) = \sum_{t \in q} IDF(t)\, f(t,d)(k1+1) / \{f(t,d)+k1[1-b+b|d|/avgdl]\}$$

The local dense baseline used word unigram and bigram TF-IDF with sublinear frequency, min_df = 2, max_df = 0.98, a 40,000-feature cap, and L2 normalization. Truncated SVD with

seven iterations projected documents and questions into 256 dimensions, followed by L2 normalization and cosine-equivalent dot-product ranking [24] . This is an LSA dense retriever rather than a pretrained neural encoder.

BM25 and LSA rankings were combined with reciprocal rank fusion [26]. The rank constant was 60. Eleven BM25 weights from 0.0 to 1.0 were tested on validation data using 0.6 Recall@10 + 0.4 FullChain@10. The selected weight was 0.70 for BM25 and 0.30 for LSA. Equal RRF remained a separate baseline.

$$RRF(q,d) = w/[60+r\_BM25(q,d)] + (1-w)/[60+r\_LSA(q,d)]$$

A standardized logistic reranker operated on the top 50 weighted-RRF candidates, following the two-stage retrieval principle in passage reranking [28] .Twelve observable features captured normalized BM25, LSA, and RRF scores; reciprocal ranks; title Jaccard overlap; IDF-weighted query coverage; source, category, and year mentions; and log document length. Training inserted an omitted gold document only into the training pair set. For answerable questions, question-level weights assigned equal total mass to positive and negative pairs; null questions contributed negative pairs only. Test reranking used the unmodified top-50 pool.

C. Sequential Evidence Selection and Dynamic Stopping

The agent converted the reranked top-30 pool into a sequential evidence order. At each hop it selected the highest utility unseen document. Utility added 0.035 for a new source and 0.015 for a new category and subtracted 0.050 times maximum title Jaccard redundancy from the reranker probability. This diversity term rewarded complementary documents without reading gold labels. The sequence contained at least two and at most eight documents.

$$u(d) = p\_rerank(d) + 0.035\ I(\text{new source}) + 0.015\ I(\text{new category}) - 0.050\ \text{redundancy}(d)$$

For every selected document, sentence scoring used IDF-weighted question coverage plus 0.12 times sentence-title overlap and a small earlier-position tie break. The title, source, date, and two best sentences were clipped to 180 regular-expression tokens. Token cost equaled question tokens, retained snippet tokens, and answer tokens. Fact coverage required the complete normalized annotated fact to remain in this clipped context.

A chain-completion classifier predicted whether a prefix contained every gold evidence document. Training covered prefix lengths two through eight using hop, score, margin, query-coverage, diversity, redundancy, source, token-cost, and null-probability features. Balanced logistic regression used $C = 1$, and validation-prefix probabilities were mapped by isotonic regression [34]. This prefix target is narrower than general tool-use trajectory success [9] and complements prior budgeted MultiHop-RAG policy evaluation [6]. The first prefix with calibrated probability at least 0.725 stopped; otherwise all eight hops were retained. Validation maximized 0.7 evidence coverage + 0.3 complete-chain rate - 0.16 normalized token cost.

D. Reader, Null Gate, and Selective Reliability

A separate null detector used character-boundary TF-IDF from three- to five-character n-grams, min_df = 2, a 30,000-feature cap, and balanced logistic regression with $C = 2$. Its validation-selected probability threshold was 0.40. When triggered, the system returned “Insufficient information.” and selected no citation.

The answer reader was trained on answerable training questions using question text plus four reranked snippets. Word unigram and bigram TF-IDF used min_df = 2, sublinear frequency, and a 40,000-feature cap. An SGD log-loss classifier used alpha = 2×10^-5, 3,000 iterations, tolerance 10^-5, class balancing, and seed 42. At inference it inspected its ten highest-probability labels and chose the first closed relational label or label present in retrieved context; otherwise it chose the highest-probability class. The same reader evaluated every retrieval policy.

Dynamic-policy reliability combined reader probability, calibrated chain-completion probability, and non-null probability. A second isotonic map was fitted to validation exact-match outcomes. Following selective-prediction principles [35], [36] and selective QA [37], the system abstained below 0.45, the maximum validation coverage with exact-match risk no greater than 0.30. The approach is consistent with evidence-calibrated refusal [4] and calibrated pairwise screening [39]; it used isotonic calibration, not the conformal alert controller studied for log escalation [68].

Reliability_raw = p_reader × p_chain × (1 - p_null)

E. Metrics and Statistical Testing

Retrieval evaluation excluded null questions and reported Recall@k, binary nDCG@k [42], MRR, and FullChain@k for k in {1,2,3,4,5,10}. End-to-end evaluation used SQuAD-style normalization, EM, and token F1 [43]. Document coverage measured selected gold URLs; fact coverage required annotated facts to survive clipping. Citation accuracy required each selected snippet to contain an annotated fact, aligning attribution with support [44]. This deterministic support rule parallels evidence guardrails in quantitative assistance [11] and ranked claim verification [50] without introducing a semantic judge.

Calibration evaluation used ten-bin expected calibration error, Brier score, and area under the risk-coverage curve. Hop ablations fixed depth from one to eight. Policy comparisons covered the dynamic agent, fixed Reranker-8, Weighted-RRF-4, Equal-RRF-4, BM25-4, and LSA-Dense-4. Paired uncertainty used 2,000 evidence-group bootstrap replicates with seed 2026. F1 and token-cost comparisons used all 511 test questions, while document-coverage comparisons used the 451 answerable questions. Differences were reported with percentile 95% intervals and two-sided empirical p-values; Holm adjustment covered twelve declared comparisons. Table III lists the fixed configuration, and Fig. 1 summarizes the agent's control flow.

TABLE III. FIXED EXPERIMENT CONFIGURATION

Component	Parameter	Value
BM25	k1 / b	1.2 / 0.75
LSA dense	Dimensions	256
RRF	Constant / BM25 weight	60 / 0.70
Reranker	Candidate depth / C	50 / 1.0
Agent	Min / max hops	2 / 8
Agent	Stop threshold	0.725

Component	Parameter	Value
Snippet	Tokens per document	180
Null gate	Threshold	0.40
Selective gate	Threshold	0.45
Bootstrap	Replicates / seed	2,000 / 2026

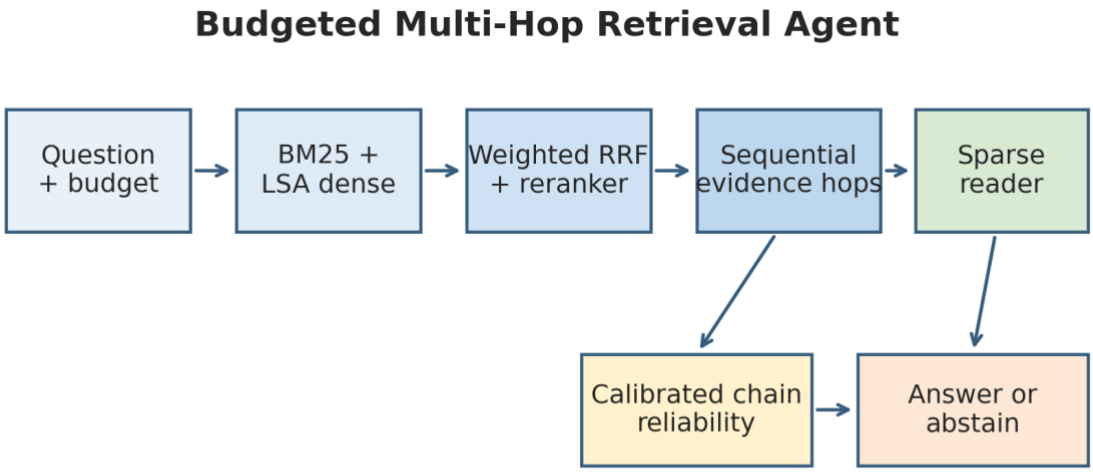

Fig. 1. Budgeted retrieval, sequential evidence selection, calibrated reliability, and answer-or-abstain control.

4. Results

A. Retrieval Quality

The held-out test set contained 451 answerable and 60 null questions. Retrieval metrics used the answerable subset; complete-system EM, F1, and cost used all 511 questions. As reported in Table IV, the reranker produced the strongest ranking results at every cutoff. At k = 4 it reached 0.7482 recall, 0.7300 nDCG, 0.4678 complete-chain retrieval, and 0.8631 MRR. Relative to BM25, the gains were 5.16, 4.19, 5.10, and 2.40 percentage points. At k = 10, reranker recall reached 0.9069, nDCG 0.8064, and complete-chain retrieval 0.7428.

TABLE IV. HELD-OUT RETRIEVAL PERFORMANCE

Method	R@4	nDCG@4	Full@4	R@10	nDCG@10	Full@10
BM25	0.6966	0.6882	0.4169	0.8679	0.7689	0.6652
LSA-Dense	0.6292	0.6008	0.3171	0.8337	0.6977	0.6275
Equal RRF	0.6783	0.6586	0.3880	0.8710	0.7502	0.6851
Weighted RRF	0.6872	0.6733	0.3991	0.8777	0.7631	0.6851

Method	R@4	nDCG@4	Full@4	R@10	nDCG@10	Full@10
Reranker	0.7482	0.7300	0.4678	0.9069	0.8064	0.7428

Fusion changes were smaller and nonuniform. At k = 4, weighted RRF improved over equal RRF from 0.6783 to 0.6872 recall, from 0.6586 to 0.6733 nDCG, and from 0.3880 to 0.3991 complete-chain retrieval. At k = 10, weighted RRF reached 0.8777 recall versus 0.8710 for equal fusion, while both complete-chain rates were 0.6851. BM25 remained stronger on several early-rank measures. The validation-selected weight therefore produced modest gains over equal fusion rather than universal dominance.

TABLE V. RETRIEVAL SLICES AT k = 4

Slice	BM25 R	BM25 Full	Rerank R	Rerank Full	N
Comparison	0.7791	0.5756	0.8391	0.6744	172
Inference	0.5992	0.2086	0.6314	0.2331	163
Temporal	0.7112	0.4741	0.7773	0.4914	116
2 documents	0.8056	0.6538	0.8739	0.7607	234
3 documents	0.6017	0.2078	0.6472	0.2078	154
4 documents	0.5238	0.0476	0.5278	0.0159	63

The slices in Table V show a sharp chain-length gradient. Reranker complete-chain retrieval at k = 4 was 0.7607 for two-document questions, 0.2078 for three-document questions, and 0.0159 for four-document questions. At k = 10 these rates increased to 0.9188, 0.6169, and 0.3968. Inference questions were also harder than comparison and temporal questions at shallow depth. Fig. 2 shows that extra cutoff depth continued to recover evidence, especially for the reranker.

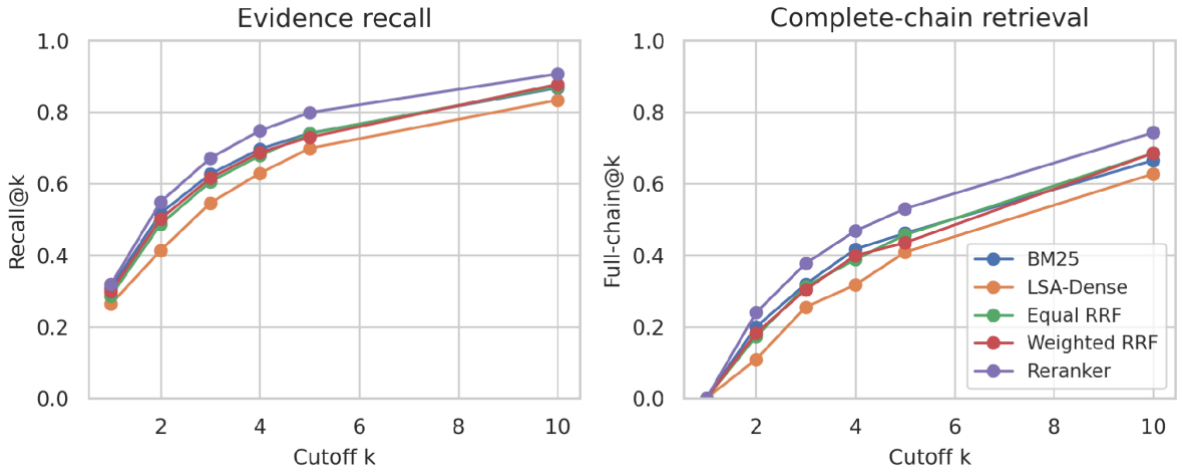

Fig. 2. Recall and complete-chain retrieval across measured cutoffs.

B. End-to-End Quality and Cost

Table VI summarizes end-to-end quality, evidence, and cost. EM, F1, and mean tokens use all 511 questions, while the evidence columns use the 451 answerable questions. Weighted-RRF-4 had the highest observed overall F1 at 0.6290 with a mean cost of 363.20 tokens. Equal-RRF-4 followed at 0.6270 and 361.77 tokens. Reranker-4 reached 0.6223 F1 despite better evidence retrieval. Under null gating, Reranker-8 selected eight documents for each answerable question, rejected the 60 null questions without retrieval, and reached 0.6243 F1 at a 666.29-token all-test mean. Improved document ranking did not translate proportionally into answer accuracy under the fixed reader and clipped snippets.

TABLE VI. END-TO-END QUALITY, EVIDENCE, AND COST

Policy	EM	F1	Doc. cov.	Fact cov.	Citation acc.	Full chain	Mean tokens
BM25-4	0.6164	0.6172	0.6966	0.3708	0.2328	0.4169	366.40
LSA-Dense-4	0.6067	0.6074	0.6292	0.3154	0.1951	0.3171	357.96
Equal-RRF-4	0.6262	0.6270	0.6783	0.3485	0.2178	0.3880	361.77
Weighted-RRF-4	0.6282	0.6290	0.6872	0.3585	0.2239	0.3991	363.20
Reranker-4	0.6223	0.6223	0.7482	0.3958	0.2489	0.4678	361.59
Reranker-8	0.6243	0.6243	0.8758	0.4497	0.1452	0.6763	666.29
Dynamic-Agent	0.6184	0.6184	0.8601	0.4418	0.2115	0.6430	544.16
Dynamic+Selective	0.3464	0.3464	0.8601	0.4418	0.2115	0.6430	544.62

The dynamic agent achieved 0.6184 F1, 0.8601 document coverage, and 0.6430 complete-chain retrieval at 544.16 tokens over 5.68 selected hops. Compared with Reranker-8, it saved 122.13 tokens and 1.38 mean hops while losing 1.57 document-coverage points, 3.33 complete-chain points, and 0.59 F1 points. Fig. 3 places it between the four-document policies and the null-gated Reranker-8 policy in cost. On the F1-token plane, four-document fusion remained above and to the left.

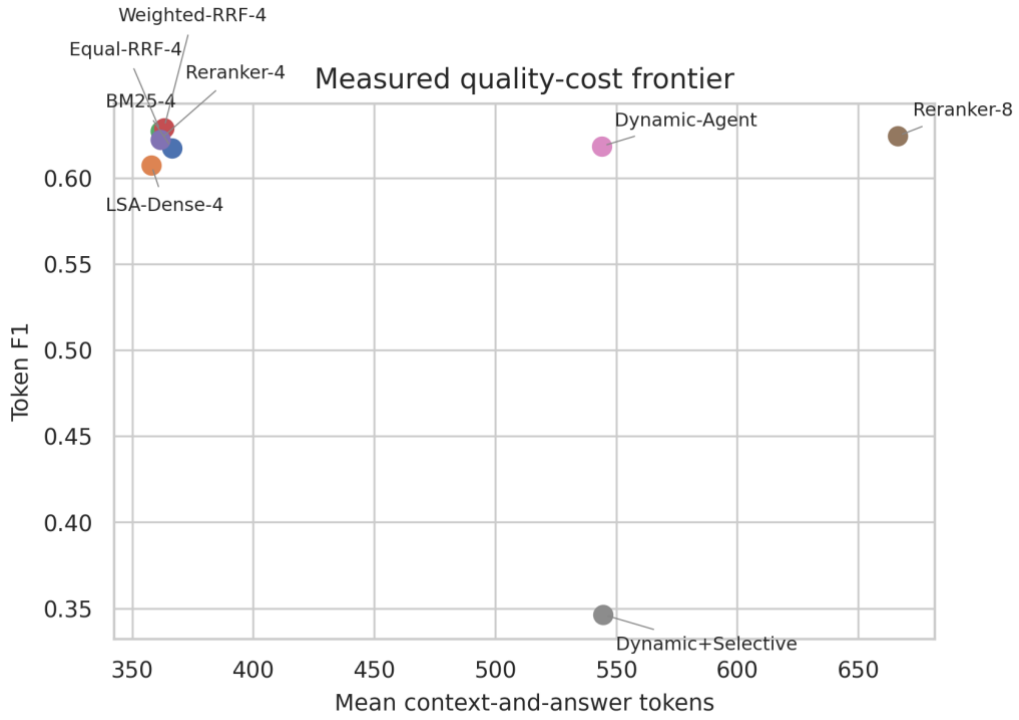

Fig. 3. Measured answer-quality and context-cost positions for all policies.

C. Question Type, Hop Depth, and Abstention

TABLE VII. DYNAMIC-AGENT RESULTS BY QUESTION TYPE

Type	N	F1	Doc. cov.	Fact cov.	Citation acc.	Mean tokens
Comparison	172	0.3837	0.9302	0.4981	0.2145	571.35
Inference	163	0.8834	0.7960	0.4790	0.2377	671.40
Temporal	116	0.3966	0.8463	0.3060	0.1703	579.08
Null	60	1.0000	—	—	—	53.03

Table VII shows the question-type breakdown; evidence metrics are undefined for null questions and are marked accordingly. Dynamic-agent F1 was 0.8834 on inference, 0.3837 on comparison, and 0.3966 on temporal questions. The null gate correctly rejected all 60 null questions and rejected none of the 451 answerable questions. The selective policy accepted 150 answerable questions, giving 0.3326 coverage and 0.2200 risk. Selection was uneven: 110 of 163 inference questions were accepted and 109 were correct, whereas only three of 26 accepted comparison questions and five of 14 accepted temporal questions were correct. Because reliability selection followed retrieval, mean cost remained 544.62 tokens.

TABLE VIII. FIXED-HOP ABLATION ON ANSWERABLE QUESTIONS

Hops	F1	Doc. coverage	Full chain	Mean tokens
1	0.5286	0.3184	0.0000	136.53
2	0.5619	0.5595	0.2506	226.23

Hops	F1	Doc. coverage	Full chain	Mean tokens
4	0.5632	0.7561	0.4745	402.38
6	0.5632	0.8337	0.5876	574.91
8	0.5765	0.8764	0.6763	747.35

Table VIII shows that evidence coverage rose regularly with hop depth, whereas answer F1 did not. For two-document questions, document coverage increased from 0.4017 at one hop to 0.9380 at eight and complete-chain retrieval reached 0.8761, while F1 changed from 0.4487 to 0.4957. Three-document coverage rose from 0.2554 to 0.8312 and four-document coverage from 0.1627 to 0.7579, but their F1 curves remained nonmonotonic. Fig. 4 separates the stable evidence gain from the irregular answer response.

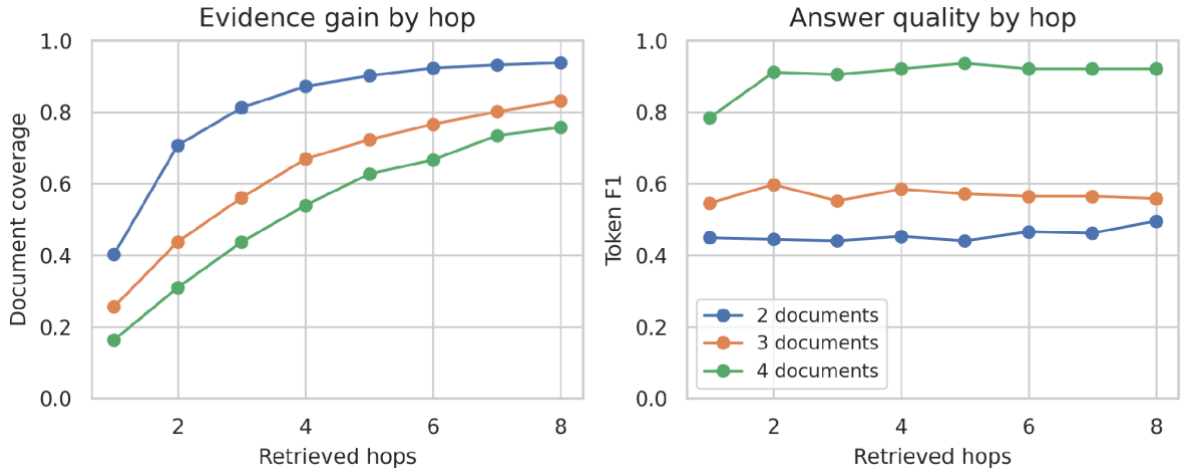

Fig. 4. Evidence coverage and answer F1 by fixed hop count and required chain size.

D. Calibration and Selective Risk

TABLE IX. ANSWER-RELIABILITY CALIBRATION

Reliability	ECE	Brier	AURC
Raw	0.2671	0.2902	0.2755
Isotonic	0.0321	0.2152	0.2717

Table IX shows that isotonic calibration reduced ECE from 0.2671 to 0.0321 and Brier score from 0.2902 to 0.2152. AURC changed from 0.2755 to 0.2717, showing that calibration corrected probability scale much more than reliability ordering. At 9.98% selective coverage, exact-match risk was 0.0444; at 19.96% it was 0.1000, at 29.93% it was 0.1926, and at full coverage it was 0.4324. Figs. 5 and 6 show both effects.

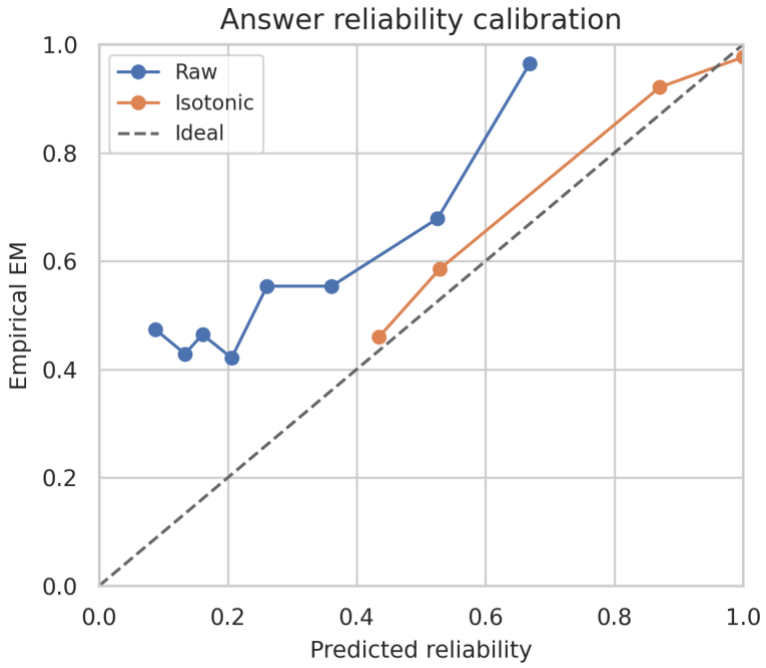

Fig. 5. Raw and isotonic answer-reliability curves on answerable test questions.

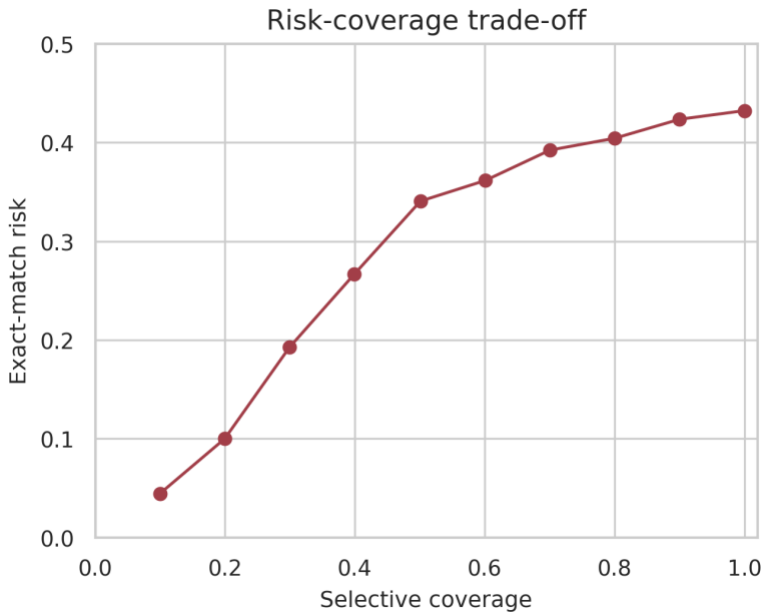

Fig. 6. Exact-match risk as calibrated selective coverage increases.

E. Ablations, Significance, and Failure Localization

Table X compares the complete dynamic policy with fixed-depth retrieval and fusion variants. Because depth, ordering, and retrieval components differ across these named policies, the table is interpreted as an operating-policy comparison; the isolated depth effect is reported in Table VIII.

TABLE X. POLICY AND COMPONENT COMPARISONS

Variant	F1	Doc. cov.	Full chain	Mean tokens	ΔF1
Dynamic policy	0.6184	0.8601	0.6430	544.16	+0.0000
Reranker, depth 8	0.6243	0.8758	0.6763	666.29	+0.0059
Weighted RRF, depth 4	0.6290	0.6872	0.3991	363.20	+0.0106
Equal RRF, depth 4	0.6270	0.6783	0.3880	361.77	+0.0086
BM25, depth 4	0.6172	0.6966	0.4169	366.40	-0.0012
LSA dense, depth 4	0.6074	0.6292	0.3171	357.96	-0.0110

The paired bootstrap in Table XI reports Dynamic-Agent minus each listed comparator. The Reranker-8 F1 difference was -0.0059, with a 95% interval of [-0.0246, 0.0115] and Holm-adjusted p = 1.0000. The token-cost difference was -122.13, with an interval of [-149.13, -96.44], and the answerable-only document-coverage difference was -0.0157, with an interval of [-0.0260, -0.0073]; their adjusted p-values were 0.0120 and 0.0120. All four F1 confidence intervals included zero, while every document-coverage and token-cost contrast remained nonzero after correction. Fig. 7 visualizes the four F1 intervals.

TABLE XI. ALL PAIRED EVIDENCE-GROUP BOOTSTRAP COMPARISONS

Comparator	Metric	Difference	95% CI	Holm p
BM25-4	F1	0.0012	[-0.0228, 0.0257]	1.0000
BM25-4	DocumentCoverage	0.1635	[0.1315, 0.1971]	0.0120
BM25-4	TokenCost	177.7632	[152.6950, 200.5848]	0.0120
Equal-RRF-4	F1	-0.0086	[-0.0352, 0.0176]	1.0000
Equal-RRF-4	DocumentCoverage	0.1818	[0.1513, 0.2167]	0.0120
Equal-RRF-4	TokenCost	182.3894	[157.3564, 205.4158]	0.0120
Reranker-4	F1	-0.0039	[-0.0269, 0.0201]	1.0000
Reranker-4	DocumentCoverage	0.1120	[0.0874, 0.1365]	0.0120
Reranker-4	TokenCost	182.5714	[158.0062, 205.4265]	0.0120
Reranker-8	F1	-0.0059	[-0.0246, 0.0115]	1.0000
Reranker-8	DocumentCoverage	-0.0157	[-0.0260, -0.0073]	0.0120
Reranker-8	TokenCost	-122.1331	[-149.1296, -96.4358]	0.0120

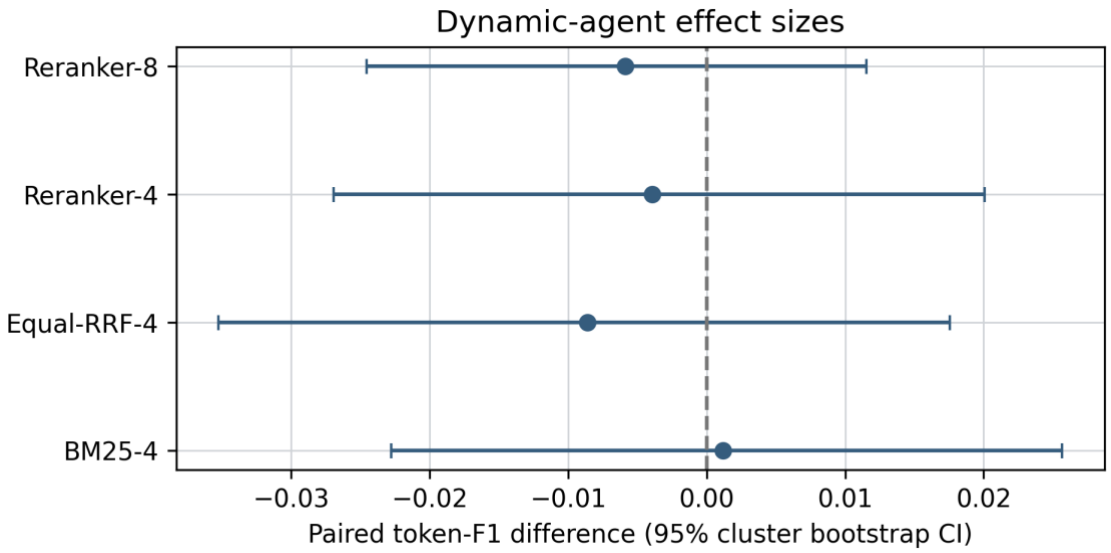

Fig. 7. Paired token-level F1 differences for the dynamic agent with 95% evidence-group bootstrap intervals.

The failure taxonomy in Table XII identifies snippet retention as the largest bottleneck. Of 511 dynamic traces, 218 selected a gold document but omitted an annotated fact from its retained snippet, and 143 had the gold documents in the candidate pool but failed to select the complete chain. Only 18 lacked a gold document in the top-50 candidate pool. Thirty-four retained all facts but answered incorrectly, and 35 returned a correct answer with at least one unsupported selected citation. Sixty were correct null rejections. Three traces satisfied all three conditions: complete retained evidence, a correct answer, and fully supported citations. The observed trace in Fig. 8 retrieved all three annotated documents and answered correctly, but it also selected five distractors, retained two of three annotated facts, and supported two of eight selected citations.

TABLE XII. DYNAMIC-AGENT FAILURE TAXONOMY

Outcome	N	Share of test
Gold document selected; fact omitted by snippet	218	42.66%
Gold document absent from candidate pool	18	3.52%
Correct rejection of null query	60	11.74%
Candidate present but not selected	143	27.98%
All facts retained; reader answer wrong	34	6.65%
Correct answer with unsupported citation	35	6.85%
Joint success	3	0.59%

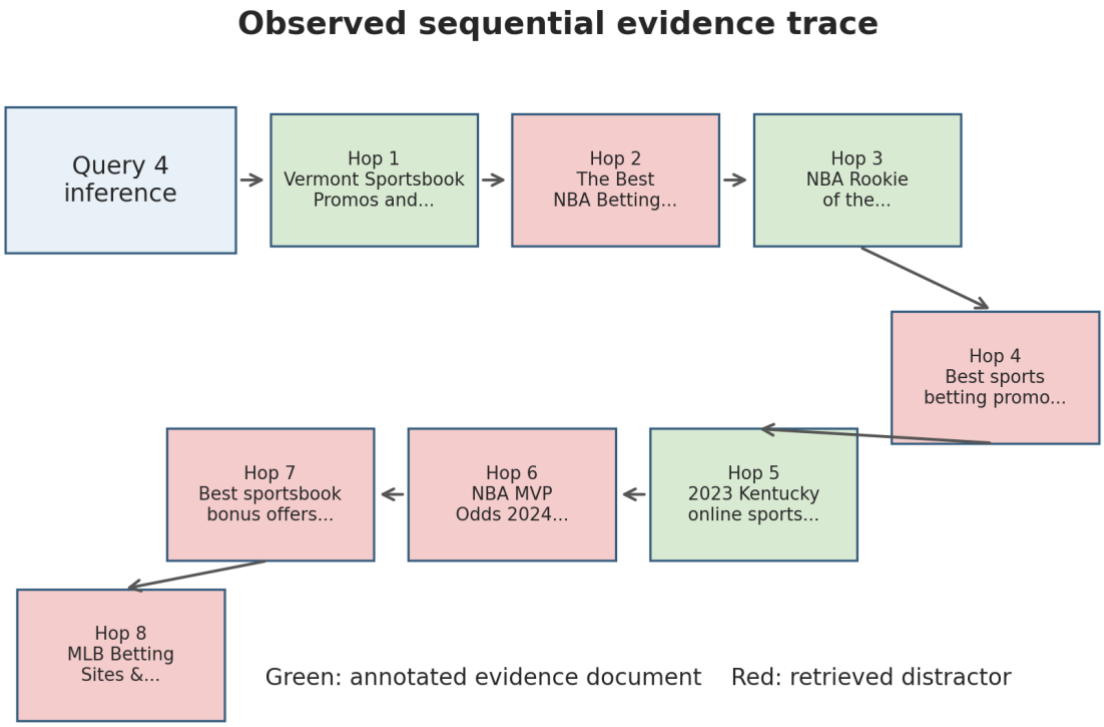

Fig. 8. Observed eight-hop trace for test question 4; three selected nodes are annotated evidence and five are distractors.

5. Discussion

The central result is a measured evidence-cost trade-off rather than a universal accuracy gain. Sequential multi-hop systems motivate conditional retrieval [17]–[20], while offline policy evaluation [83] and cost-sensitive admission control [79] emphasize comparison at an operating point. Dynamic selection raised document coverage by 16.35 points over BM25-4 and 11.20 points over Reranker-4. Relative to null-gated Reranker-8, it removed 122.13 mean tokens with a confidence interval for F1 that included zero, while accepting a small but significant coverage loss. It is therefore a cheaper approximation to deep reranking when evidence coverage matters, not the highest-F1 policy.

The gap between retrieval and answering is consequential. Reranking improved recall, nDCG, and complete-chain recovery, yet Weighted-RRF-4 produced the highest observed F1. Greater hop depth reliably improved document coverage but yielded irregular F1. Adaptive retrieval [30], [31], [7], [32] does not by itself ensure that retained spans contain usable facts. Retrieval-quality prediction [21], grounded HDFS narratives [65], and incident evidence cards [67] point to the same post-retrieval conversion problem. A stopping objective must reflect retained support, reader use, and citation validity, not document acquisition alone.

Fusion likewise did not guarantee improvement. The 256-dimensional LSA retriever was weaker than BM25, so validation assigned 70% weight to BM25. Weighted RRF improved several aggregate measures over equal fusion, but the advantage reversed at some cutoffs. This result is consistent with cross-domain retrieval variation [25] and with the need for learned hybrid reranking on FiQA [29]; spectral rank-aggregation theory [27] does not imply that one

empirical fusion weight dominates across datasets. The study therefore makes no claim about a pretrained neural dense model.

Calibration sharply improved probability accuracy while exposing a subgroup problem. Isotonic mapping reduced ECE and Brier score, consistent with calibration and selective prediction [34], [37], but AURC barely changed because rescaling did not substantially reorder hard cases. Comparable risk-communication concerns appear in calibrated fusion [38], recruiter refusal [39], and medical safety cards [54]. The global threshold achieved 0.220 risk at 0.333 coverage but performed far better for inference than relational questions. Type-conditioned calibration or minimum fact-support constraints are therefore more defensible than a single universal threshold.

Evidence diagnostics place span retention and citation precision ahead of candidate generation. Only 3.52% of traces lacked a gold document in the top-50 pool, whereas 42.66% selected a gold document but clipped an annotated fact. Citation accuracy also fell as distractors entered the selected set. Citation and component-level evaluations [44]–[46], evidence-chain provenance [3], accounting retrieval [47], and scientific evidence displays [10] all support separating a correct answer from attributable support. The metric used here is deliberately strict and gives no credit to paraphrases without the annotated fact.

The study uses one 609-document news corpus, one grouped split, an LSA retriever, a feature-based reranker, and a sparse supervised reader. It does not determine how a neural dense retriever, cross-encoder, or generative model behaves under the same controller. Regex token count is a transparent workload measure rather than latency, energy, or billed API cost; hybrid-cloud deployment [78], GPU demand forecasting [76], and power-aware inventory planning [80] include additional resource dimensions. Domain-bridge evidence [15] also cautions against treating the grouped split as robustness under temporal or domain shift.

These results support two operating choices. Weighted-RRF-4 is preferable for the highest observed F1 at low token cost. Dynamic retrieval is preferable when higher evidence coverage is needed but Reranker-8 is too expensive. The next improvement should rerank sentences or evidence spans, make stopping sensitive to retained facts and citation support, and calibrate by question family. Scientific search cards [10], accounting disclosure cards [48], and prompt-injection risk cards [75] show how those signals could be presented for human review. These are interface implications, not claims tested by the current experiment.

6. Conclusion

A complete MultiHop-RAG evaluation showed that retrieval depth, evidence reliability, answer quality, and cost are related but not interchangeable. Supervised reranking produced the strongest evidence ranking, reaching 0.9069 Recall@10 and 0.7428 complete-chain retrieval. Relative to the null-gated Reranker-8 policy, calibrated dynamic stopping reduced mean context-and-answer cost from 666.29 to 544.16 tokens, while the 95% confidence interval for the paired F1 difference included zero. Isotonic calibration reduced ECE to 0.0321, and selective abstention lowered risk at the cost of substantial coverage and uneven performance across question types. The dominant failures occurred after document retrieval: relevant facts were clipped from snippets or extra citations lacked annotated support. A reliable multi-hop

agent therefore needs span-level evidence control and citation-aware stopping in addition to stronger document retrieval.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Author's contribution using CRediT roles.

References

[1] W. Pi, "Efficient Information Retrieval and Response Generation with Retrieval-Augmented Generation (RAG)," 2024. doi: 10.59350/r9dj1-zkx52.

[2] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks," *Advances in Neural Information Processing Systems 36*, pp. 48573-48602, 2023. doi: 10.52202/075280-2109.

[3] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 520-533, 2025. doi: 10.51903/jtie.v4i2.535.

[4] 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," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 502-520, 2025. doi: 10.51903/jtie.v4i2.536.

[5] Y. Chen, and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," *International Journal of Graphic Design*, vol. 4, no. 1, pp. 141-164, 2026. doi: 10.51903/ijgd.v4i1.3552.

[6] 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," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 649-662, 2025. doi: 10.51903/jtie.v4i3.543.

[7] 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," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036-7050, 2024. doi: 10.18653/v1/2024.naacl-long.389.

10.55415/deep-2024-0007.v1[8] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance," arXiv:2305.05176, 2023, doi: 10.48550/arXiv.2305.05176.

[9] Ziliang Samuel Zhong, Chenyu Li, and Hengning Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 341-360, 2026. doi: 10.51903/jtie.v5i1.539.

[10] 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, 2025, doi: 10.51903/ijgd.v3i2.3698.

[11] Z. Li, Kai Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant:

A Compact Reproducible Benchmark Using SEC and FRED Data," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 75-90, 2026. doi: 10.51903/jtie.v5i2.541.

[12] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, et al., "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369-2380, 2018. doi: 10.18653/v1/d18-1259.

[13] X. Ho, A. K. Duong Nguyen, S. Sugawara, and A. Aizawa, "Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609-6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

[14] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "♫ MuSiQue: Multihop Questions via Single-hop Question Composition," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 539-554, 2022. doi: 10.1162/tacl_a_00475.

[15] 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, 2025, pp. 559–567.

[16] Y. Zhang, P. Nie, A. Ramamurthy, and L. Song, "Answering Any-hop Open-domain Questions with Iterative Document Reranking," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 481-490, 2021. doi: 10.1145/3404835.3462853.

[17] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014-10037, 2023. doi: 10.18653/v1/2023.acl-long.557.

[18] Z. Zhuang, Z. Zhang, S. Cheng, F. Yang, J. Liu, S. Huang, et al., "EfficientRAG: Efficient Retriever for Multi-Hop Question Answering," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3392-3411, 2024. doi: 10.18653/v1/2024.emnlp-main.199.

[19] H. Lee, S. Yang, H. Oh, and M. Seo, "Generative Multi-hop Retrieval," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1417-1436, 2022. doi: 10.18653/v1/2022.emnlp-main.92.

[20] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9248-9274, 2023. doi: 10.18653/v1/2023.findings-emnlp.620.

[21] 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," 2025. doi: 10.48550/arxiv.2511.19481.

[22] S. Robertson, and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, vol. 4, no. 1-2, pp. 1-174, 2009. doi: 10.1561/1500000019.

[23] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, et al., "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769-6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[24] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990. doi: 10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9.

10.63317/3tsv3vgn8kns[25] N. Thakur et al., "BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models," in Proc. Neural Information Processing Systems Track on Datasets and Benchmarks, 2021, pp. 1–18, doi: 10.48550/arXiv.2104.08663.

[26] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet

and individual rank learning methods," *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758-759, 2009. doi: 10.1145/1571941.1572114.

[27] Z. Samuel Zhong, and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," *Information and Inference: A Journal of the IMA*, vol. 13, no. 3, 2024. doi: 10.1093/imaiai/iaae020.

10.6028/nist.sp.1250.deep-udel_fang[28] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv:1901.04085, 2019, doi: 10.48550/arXiv.1901.04085.

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

10.3102/ip.24.2106302[30] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[31] Z. Jiang, F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, et al., "Active Retrieval Augmented Generation," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969-7992, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

[32] 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," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12991-13013, 2024. doi: 10.18653/v1/2024.acl-long.702.

[33] 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," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 91-103, 2026. doi: 10.51903/jtie.v5i2.538.

10.5220/0006146800700077[34] 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.

[35] Y. Wiener, and R. El-Yaniv, "Agnostic Pointwise-Competitive Selective Classification," *Journal of Artificial Intelligence Research*, vol. 52, pp. 171-201, 2015. doi: 10.1613/jair.4439.

10.52202/075280-0996[36] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Proc. Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4878–4887.

[37] A. Kamath, R. Jia, and P. Liang, "Selective Question Answering under Domain Shift," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684-5696, 2020. doi: 10.18653/v1/2020.acl-main.503.

[38] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 215-238, 2026. doi: 10.51903/jtie.v4i1.485.

[39] J. Jin, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 625-648, 2025. doi: 10.51903/jtie.v4i3.529.

[40] Z. S. Zhong, and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," 2024. doi: 10.48550/arxiv.2408.05944.

[41] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," *J-INTECH*, vol. 14, no. 02, pp. 215-231, 2026. doi: 10.32664/j-intech.v14i02.2228.

[42] K. Järvelin, and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422-446, 2002. doi:

10.1145/582415.582418.

[43] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383-2392, 2016. doi: 10.18653/v1/d16-1264.

[44] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling Large Language Models to Generate Text with Citations," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465-6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

[45] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150-158, 2024. doi: 10.18653/v1/2024.eacl-demo.16.

[46] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338-354, 2024. doi: 10.18653/v1/2024.naacl-long.20.

[47] Y. Chen, S. Zhou, and E. Lin, "Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 649-663, 2025. doi: 10.51903/jtie.v4i3.542.

[48] 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, 2025, doi: 10.51903/ijgd.v3i2.3710.

[49] S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 60-74, 2026. doi: 10.51903/jtie.v5i2.544.

[50] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 327-340, 2026. doi: 10.51903/jtie.v5i1.549.

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

[52] Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–210, 2025, doi: 10.51903/ijgd.v3i1.3715.

[53] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415–436, 2025, doi: 10.51903/ijgd.v3i2.3616.

[54] B. Zhou, H. Wang, and X. Chang, "Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, 2025, doi: 10.51903/ijgd.v3i2.3696.

[55] C. Li, B. Zhou, and K. Gao, "Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–394, 2025, doi: 10.51903/ijgd.v3i2.3709.

[56] 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, 2025, doi: 10.51903/ijgd.v3i2.3697.

[57] Y. Li, S. Lu, and L. Zhao, "LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–215, 2025, doi: 10.51903/ijgd.v3i1.3661.

[58] X. Chang, Y. Lu, and Z. S. Zhong, "Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews," *JEECS (Journal of Electrical Engineering and Computer Sciences)*, vol. 11, no. 1, pp. 9-22, 2026. doi: 10.54732/jeecs.v11i1.2.

[59] Yifan Zhang, and Hailey Zhang, "A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 464-486, 2025. doi: 10.51903/jtie.v4i2.547.

[60] 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.

[61] Y. Zhang, and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," *International Journal of Graphic Design*, vol. 3, no. 1, pp. 214-229, 2025. doi: 10.51903/ijgd.v3i1.3722.

[62] Q. Xin, "Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education," J. Artificial Intelligence Educ., vol. 2, no. 1, 2026, doi: 10.66053/jaaie.v2i1.348.

[63] H. Xu, Y. Chen, and A. Med, "Automatic Detection and Explanation of Dark Patterns from Interface Microcopy: Empirical Comparison of BERT-Style Encoders, RoBERTa-Style Encoders, and LLM-Style Decoders on the ec-darkpattern Dataset," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 590-612, 2025. doi: 10.51903/jtie.v4i3.491.

[64] Q. Xin, "Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction," Interdiscip. J. Pedagog. Res. Media Technol., vol. 2, no. 1, 2026, doi: 10.64268/inspire.v2i1.117.

[65] X. Sun, Ziliang Samuel Zhong, and Q. Wu, "Retrieval-Grounded HDFS Log Anomaly Detection and Deterministic Failure Narrative Generation," *Journal of Computational Systems and Applications*, vol. 3, no. 1, pp. 15-30, 2026. doi: 10.64229/j6d7fr94.

[66] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," *Emerging Information Science and Technology*, vol. 6, no. 2, pp. 125-146, 2025. doi: 10.18196/eist.v6i2.31232.

[67] J. Nie, G. Liu, and L. Zhao, "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," *International Journal of Graphic Design*, vol. 4, no. 1, pp. 179-185, 2026. doi: 10.51903/ijgd.v4i1.3703.

[68] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC)*, vol. 8, no. 2, pp. 247, 2026. doi: 10.28989/avitec.v8i2.3974.

[69] J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28–43, 2026, doi: 10.33474/infotron.v6i1.24923.

[70] Q. Xin, "Self-Supervised Log Anomaly Detection with LogBERT-Style Transformers: Full Empirical Evaluation on a Reproducible SynHDFS Benchmark," *JEECS (Journal of Electrical Engineering and Computer Sciences)*, vol. 11, no. 1, pp. 23-35, 2026. doi: 10.54732/jeecs.v11i1.3.

[71] Y. Li, and S. Lu, "Language-Guided Feature Selection for DDoS and Intrusion Detection on CICIDS2017," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 284-305, 2025. doi: 10.51903/jtie.v4i1.531.

[72] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," *Applied and Computational Engineering*, vol. 69, no. 1, pp. 64-70, 2024. doi: 10.54254/2755-2721/69/20241511.

[73] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," *Applied and Computational Engineering*, vol. 64, no. 1, pp. 94-99, 2024. doi: 10.54254/2755-2721/64/20241350.

[74] Q. Xin, "Host-Based Intrusion Detection with System Call Sequences: Window Localization and Forensic Narratives," *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC)*, vol. 8, no. 2, pp. 325, 2026. doi: 10.28989/avitec.v8i2.3973.

[75] 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," *International Journal of Graphic Design*, vol. 4, no. 1, pp. 186-191, 2026. doi: 10.51903/ijgd.v4i1.3699.

[76] S. Zhao, J. Bai, and D. Roberson, "Multi-Horizon GPU Demand Forecasting with Workload Semantics and Operational Risk Curves: An Empirical Study on Alibaba Clusterdata GPU Trace," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 544-571, 2025. doi: 10.51903/jtie.v4i3.498.

[77] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 306-324, 2025. doi: 10.51903/jtie.v4i1.546.

[78] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," *Journal of Information Systems and Informatics*, vol. 7, no. 3, pp. 2182-2195, 2025. doi: 10.51519/journalisi.v7i3.1170.

[79] Siming Zhao, Y. Ren, and Xiaohan Chang, "Profit-Aware Spot GPU Admission Control with Cost-Sensitive Loss and Evidence-Grounded Policy Memos for AI Workload Supply-Demand Matching," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 45-59, 2026. doi: 10.51903/jtie.v5i2.545.

[80] S. He, J. Nie, and C. Li, "Power-Aware Inventory Planning for AI Infrastructure Using Job-Level Forecasting and LLM Workload Explanations," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 341-359, 2026. doi: 10.51903/jtie.v5i1.548.

[81] 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," *International Journal of Graphic Design*, vol. 3, no. 1, pp. 196-213, 2025. doi: 10.51903/ijgd.v3i1.3723.

[82] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU Memory Requirement Prediction for Deep Learning Task Based on Bidirectional Gated Recurrent Unit Optimization Transformer," *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV)*, pp. 31-35, 2025. doi: 10.1109/aivrv67401.2025.11350369.

[83] J. Mu, T. Ye, and P. Patel, "Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 521–543, 2025, doi: 10.51903/jtie.v4i3.500.

[84] J. Bai, H. Wang, Q. Wu, and B. Zhang, "Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 17-38, 2026. doi: 10.51903/jtie.v5i1.468.

[85] T. Ye, J. Mu, and J. Hunter, "Off-Policy Evaluation and Conservative Policy Selection for Slot-Level Dynamic Bidding and Ranking on the Open Bandit Dataset (Small)," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 178-199, 2026. doi: 10.51903/jtie.v5i1.503.

[86] Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 493-520, 2025. doi: 10.51903/jtie.v4i3.466.

[87] J. Bai and Q. Wu, "Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study," Int. J. Electron. Commun. Syst., vol. 6, no. 1, 2026, doi: 10.24042/ijecs.v6i1.30533.

[88] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," *International Journal of Graphic Design*, vol. 3, no. 1, pp. 216-240, 2025. doi: 10.51903/ijgd.v3i1.3713.

[89] H. Tu, S. Zhao, and A. Zhou, "Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 210–226, 2025, doi: 10.51903/ijgd.v3i1.3714.

[90] J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards,"

International Journal of Graphic Design, vol. 4, no. 1, pp. 192-208, 2026. doi: 10.51903/ijgd.v4i1.3702.