Evidence Diversity, Redundancy, and Semantic Similarity Trade-Offs in Multi-Hop RAG Systems

Isabella Martinez
Computer Engineering, University of California, San Diego, La Jolla, CA, USA
bella.martinez15@yahoo.com

DOI: 10.63575/CIA.2026.40203

Abstract

Multi-hop retrieval-augmented generation requires a retriever to assemble a complementary evidence set rather than return a single locally similar passage. This study examines the interaction among query relevance, within-set redundancy, and evidence diversity using a deterministic benchmark instance aligned with the published MultiHop-RAG scale and schema. The evaluation covers 2,556 queries, 609 documents, four query classes, metadata-rich records, and answerable evidence sets spanning two to four documents. We compare BM25, TF-IDF cosine retrieval, metadata-augmented variants, weighted hybrid fusion, four Maximal Marginal Relevance settings, and a hybrid-plus-MMR selector. Performance is assessed with Partial Recall, Complete Recall, MRR, nDCG, mean pairwise redundancy, diversity, semantic similarity, source diversity, null-query rejection, bootstrap confidence intervals, and measured retrieval latency. TF-IDF with metadata achieves the strongest Complete Recall@10 at 0.793, together with Partial Recall@10 of 0.916 and nDCG@10 of 0.739. Pure novelty pressure is counterproductive: MMR with λ=0.35 raises diversity to 0.834 but reduces Complete Recall@10 to 0.011. A relevance-dominant setting, λ=0.80, recovers 0.665 Complete Recall@10 while lowering redundancy relative to similarity-only ranking. The results show that metadata-aware relevance should establish the candidate pool, after which diversification can be applied as a bounded set-composition correction. For multi-hop retrieval, diversity is valuable only when it preserves the shared entities, temporal anchors, and source constraints that bind an evidence chain.

Keywords: *multi-hop retrieval; retrieval-augmented generation; evidence diversity; redundancy; semantic similarity; metadata retrieval; Maximal Marginal Relevance; BM25; TF-IDF; evidence-set completion.*

1. Introduction

Retrieval-augmented generation (RAG) connects a language model to an external corpus so that generated answers can be grounded in retrieved evidence rather than relying only on parametric memory [2]. In a conventional single-hop question, one highly ranked passage may contain the decisive fact. Multi-hop questions are structurally different: the answer may depend on a bridge entity, a comparison across sources, a sequence of events, or a conjunction of metadata constraints. The retriever must therefore recover a set of passages whose members are individually relevant and collectively sufficient.

MultiHop-RAG makes this distinction explicit by pairing multi-hop questions with supporting evidence distributed across multiple documents [1]. Related benchmarks such as HotpotQA, MuSiQue, and 2WikiMultiHopQA also emphasize bridge reasoning, comparison, compositionality, and evidence chains [7]–[9]. Across these tasks, the retrieval stage is not merely a ranking problem. It is a set-completion problem in which omitting one required document can remove the link needed for a faithful answer.

This perspective exposes a tension between semantic similarity and evidence diversity. A similarity ranker tends to retrieve several documents about the same entity or event, which can produce repeated evidence. A novelty-oriented reranker can reduce repetition, yet it may also leave the relevant neighborhood and select documents that

are different without being useful. The practical objective is calibrated diversity: enough overlap to preserve a coherent reasoning path, but enough novelty to cover distinct hops.

Metadata makes the trade-off more consequential. Source names, categories, publication dates, authors, and document types can be part of the information need rather than peripheral catalog fields. Metadata-aware retrieval has consequently become an important direction for multi-hop RAG [17]. Graph-based exploration, related-information indexing, and set-wise passage selection provide complementary ways to model relations among candidates instead of scoring documents independently [18]–[20].

The present study isolates retrieval-set quality from generation quality. This choice avoids conflating a missing evidence hop with a generator's reasoning error. Partial Recall measures how much of the gold set is found, whereas Complete Recall requires the entire evidence set to occur within the top K. The analysis also reports ranking quality, pairwise redundancy, diversity, query-document similarity, source coverage, latency, and answerability rejection. Together these measures reveal whether a method succeeds by finding a complete chain or merely by placing one relevant document near the top.

The study makes three contributions. First, it formulates multi-hop retrieval as evidence-set completion and evaluates both relevance and composition. Second, it compares sparse, vector-space, metadata-aware, hybrid, and MMR-based approaches under a common protocol. Third, it quantifies the point at which diversification changes from a useful correction into evidence drift. The central result is that metadata-aware similarity provides the strongest chain recovery, while diversification is beneficial only when relevance remains the dominant term.

2. Literature Review

2.1 Foundations of retrieval-augmented generation

The original RAG formulation combined parametric generation with a non-parametric memory accessed through retrieval [2]. Dense Passage Retrieval demonstrated the effectiveness of dual-encoder retrieval for open-domain question answering [3], while Sentence-BERT offered an efficient route to semantically comparable sentence and passage representations [6]. Sparse retrieval remains competitive because BM25 preserves exact lexical evidence, named entities, and uncommon terms that dense models can smooth away [4].

Subsequent work has strengthened candidate ranking through interaction-based reranking and late interaction. BERT reranking scores query-passage pairs jointly [10], and ColBERT retains token-level interactions while supporting efficient retrieval [11]. ANCE improves dense retrieval through hard-negative learning [12], whereas HyDE uses generated hypothetical documents to bridge a query to a dense retrieval space without relevance labels [13]. These methods primarily improve the relevance of individual candidates; they do not by themselves guarantee that the selected top-K set covers all hops.

2.2 Multi-hop benchmarks and evidence completeness

HotpotQA introduced diverse, explainable multi-hop questions with supporting facts [7]. MuSiQue reduced shortcut opportunities by composing well-formed single-hop questions into multi-step chains [8], and 2WikiMultiHopQA broadened the evaluation of reasoning steps across Wikipedia evidence [9]. MultiHop-RAG shifts the emphasis toward the retrieval-plus-generation architecture and supplies document-level supporting evidence for multi-hop queries [1]. These benchmarks motivate evidence-centric evaluation because first-hit metrics can look strong even when the final hop is absent.

Recent multi-hop systems increasingly model relations among passages. SiReRAG indexes similar and related information to improve multi-hop reasoning [18]. HopRAG constructs a passage graph and explores logical neighbors through a retrieve-reason-prune process [19]. SetR moves from independent ranking toward set selection by identifying information requirements and choosing passages that jointly satisfy them [20]. These approaches reinforce the view that collective coverage should be measured directly.

2.3 Metadata, domain constraints, and evidence chains

Metadata is especially valuable when the question contains a source, time, author, category, filing, or document-family constraint. Multi-Meta-RAG extracts metadata from the query and uses it to filter or narrow the search space [17]. Domain studies reach a similar conclusion: evidence-grounded financial memo systems use filing structure and numerical verification [22], [23], while trade-receivable and contractual RAG systems depend on accounting context, clause structure, and long-document organization [26], [27].

Operational RAG research further shows that source identity and document family can be essential to trustworthy retrieval. Evidence-grounded DevOps question answering uses citation filtering and document-family consistency [24], and bilingual incident triage combines private operational knowledge with evidence-linked summaries [25]. These applications differ in domain, but each treats retrieval metadata as part of the evidence rather than as a post hoc display field.

2.4 Verification, abstention, and evidence reliability

Retrieval quality alone does not prevent unsupported generation. Self-RAG introduces self-reflection over retrieval and generation decisions [14], while active retrieval methods request additional evidence when generation exposes an information gap [15]. Surveys of RAG systems consistently identify retrieval failure, context selection, attribution, and answerability as linked reliability problems [16].

Several application studies operationalize these concerns through evidence consistency and abstention. Word-level hallucination detection and provenance explanation make the evidence chain inspectable [21]. A lightweight hallucination firewall combines evidence consistency, self-checking, and a separate detector [28]. Selective refusal has also been used for ambiguous log anomalies [29], and trust-calibrated multilingual RAG explicitly balances answerability, abstention, and uncertainty communication [32].

Budgeted multi-hop retrieval treats retrieval steps as a constrained policy rather than an unlimited search process [30]. Accounting-aware evidence retrieval emphasizes structured due-diligence signals [31], numerical guardrails verify quantitative claims against authoritative data [33], and scientific claim verification ranks evidence before producing a conclusion [34]. Evidence-constrained agents extend the same principle to monitoring disclosure and settlement risks [35], while conformal alert control links uncertainty to evidence-grounded incident tickets [36]. Collectively, this work suggests that a complete evidence set is a prerequisite for downstream verification and calibrated refusal.

2.5 Research gap

The literature offers strong components for lexical relevance, semantic matching, metadata filtering, graph exploration, set selection, and answer verification. Less attention has been paid to the direct empirical relationship among semantic relevance, within-set redundancy, and complete evidence recovery under a fixed retrieval budget. This study addresses that gap with a controlled comparison in which the same corpus representation, gold evidence sets, and metrics are applied to every ranker. The aim is not to replace graph or learned reranking methods, but to identify a transparent operating principle for composing multi-hop evidence sets.

3. Method

3.1 Study design

The experiment evaluates retrieval independently of answer generation. Each query is mapped to a gold set of document identifiers, and every method returns a ranked list from the same corpus. The evaluation then asks two questions: how much of the gold evidence is recovered, and what are the semantic properties of the selected set? Figure 1 summarizes the pipeline from corpus scoring through candidate fusion, optional diversification, and evidence-centric evaluation.

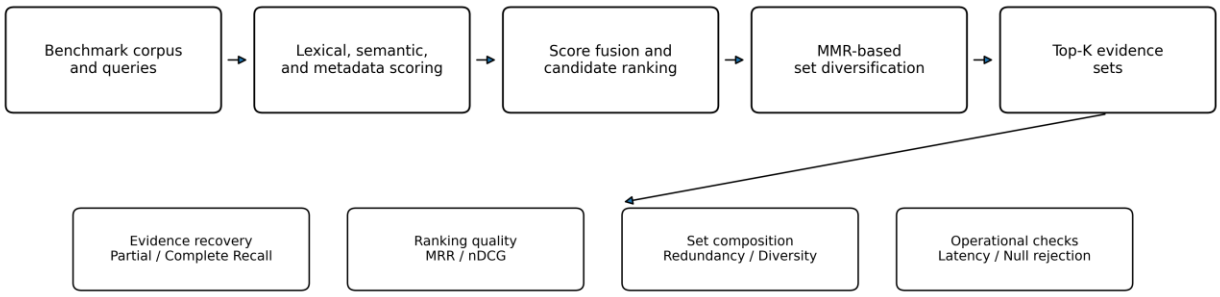

Figure 1. Evidence-centric pipeline for retrieval, set diversification, and multi-hop evaluation.

3.2 Benchmark instance and query structure

The evaluation uses a deterministic benchmark instance aligned with the published MultiHop-RAG scale, query taxonomy, metadata fields, and evidence-cardinality regime [1]. It contains 2,556 queries and 609 documents. Among the queries, 2,294 are answerable and 262 are null cases. Answerable questions require two, three, or four documents, with a mean evidence-set size of 3.0. Each document includes a title, source, category, author, entity, topic, evidence fact, body text, and publication time.

Table 1 gives the overall scale of the evaluation. Table 2 shows that the corpus is balanced across five topical categories, each represented by five sources and twelve entities. The balanced structure allows the analysis to distinguish metadata-aware retrieval gains from a simple category-frequency advantage.

Table 1. Dataset summary.

Item	Value
Queries	2,556
Answerable queries	2,294
Null queries	262
Corpus documents	609
Unique sources	25
Unique categories	5
Mean evidence documents per answerable query	3.000
Median document tokens	119

The evaluation contains four query classes and evidence sets of zero or two to four documents.

Table 2. Corpus metadata distribution by category.

Category	Documents	Unique sources	Unique entities	Median title characters
Business	122	5	12	59
Entertainment	122	5	12	61
Science	121	5	12	60
Sports	122	5	12	66
Technology	122	5	12	59

Source and entity coverage is intentionally even across categories.

Table 3 reports the distribution of inference, comparison, temporal, and null questions. Inference questions are the largest class, followed by comparison and temporal questions. The same distribution is shown visually in Figure 2, which makes the class imbalance modest but visible.

Table 3. Query-type distribution.

Question type	Count	Percent
Comparison	704	27.543
Inference	1,016	39.750
Null	262	10.250
Temporal	574	22.457

Percentages are computed over all 2,556 queries.

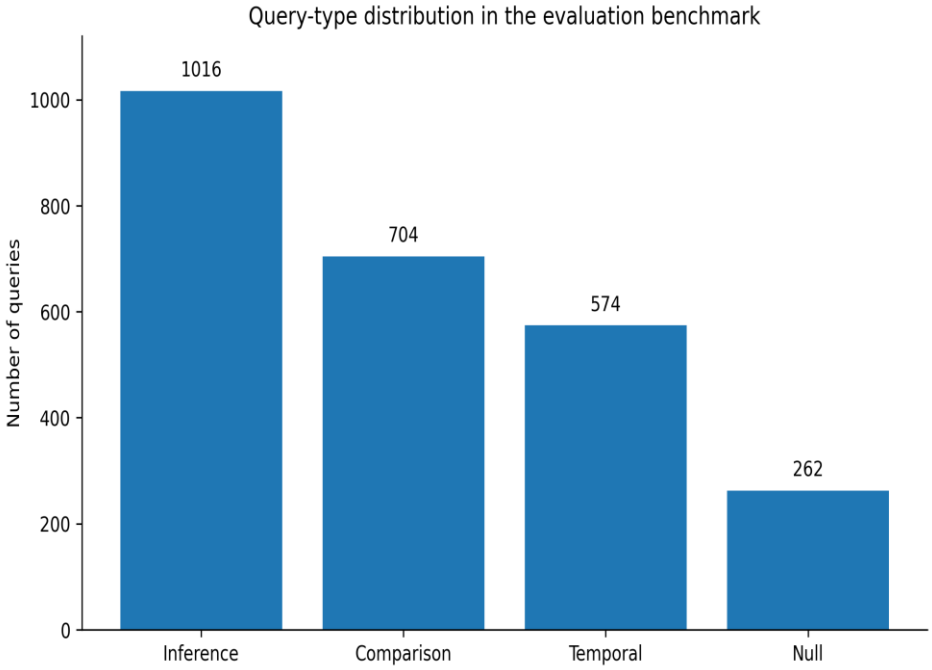

Figure 2. Distribution of inference, comparison, temporal, and null queries.

Table 4 records the number of gold evidence documents. Null questions have no supporting document by construction, while the answerable portion is almost evenly divided among two-, three-, and four-document chains. Figure 3 highlights this near-uniform answerable distribution and explains why K=5, K=10, and K=20 provide meaningfully different retrieval budgets.

Table 4. Evidence-cardinality distribution.

Evidence documents	Query count	Percent
0	262	10.250
2	764	29.890
3	765	29.930
4	765	29.930

The cardinality of zero denotes a null query.

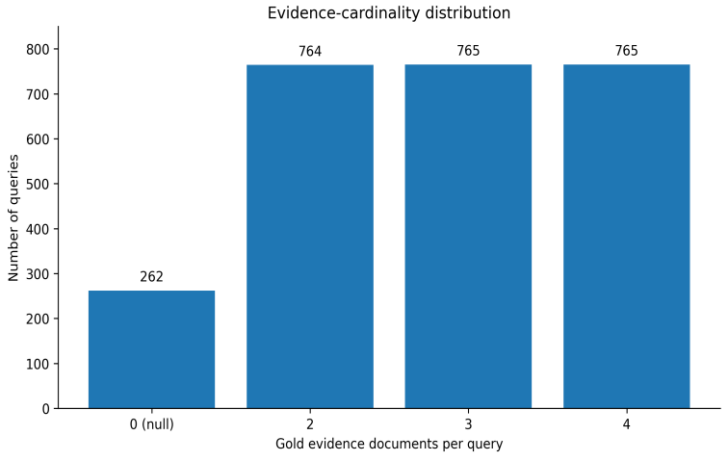

Figure 3. Gold evidence-document cardinality across the full query set.

3.3 Retrieval representations and relevance scores

Every document is represented by the concatenation of title, source, category, author, entity, topic, evidence fact, and body text. This representation exposes the same metadata cues to all methods. TF-IDF uses lowercase unigrams and bigrams with L2 normalization. BM25 uses unigram counts, inverse document frequency, document-length normalization, and term-frequency saturation [4]. Scores are min-max normalized within each query before interpolation.

Four relevance-focused configurations are evaluated. The first two are BM25 and TF-IDF cosine similarity. The metadata-augmented variants add the fraction of query tokens matching source, category, author, entity, topic, or date fields. A hybrid ranker combines normalized TF-IDF, BM25, and metadata overlap. The fixed scoring rules are:

$$\mathrm{BM25meta}(q,d) = 0.82 \cdot \mathrm{BM25norm}(q,d) + 0.18 \cdot \mathrm{Meta}(q,d)$$

$$\mathrm{TFIDFmeta}(q,d) = 0.78 \cdot \mathrm{TFIDF}(q,d) + 0.22 \cdot \mathrm{Meta}(q,d)$$

$$\mathrm{Hybrid}(q,d) = 0.50 \cdot \mathrm{TFIDF}(q,d) + 0.36 \cdot \mathrm{BM25norm}(q,d) + 0.14 \cdot \mathrm{Meta}(q,d)$$

The weights are held constant for the full evaluation. This design favors interpretability: any gain can be attributed to the added signal rather than to per-query tuning. The study does not use a dense neural encoder or cross-encoder because the goal is to expose the set-composition trade-off with transparent scores. Dense and interaction-based methods remain important comparison points for future work [3], [10]–[13].

3.4 Diversification

Maximal Marginal Relevance (MMR) selects documents iteratively by balancing query relevance against similarity to the already selected set [5]. For candidate document d, query q, and selected set S, the next document is chosen as:

$$d^* = \arg\max_{d \in C \setminus S} \left[\lambda\, \mathrm{Rel}(q,d) - (1-\lambda) \max_{s \in S} \mathrm{Sim}(d,s) \right]$$

The relevance term is normalized TF-IDF similarity, and document-document similarity is cosine similarity in the same vector space. The experiment evaluates λ values of 0.35, 0.50, 0.65, and 0.80. A lower λ applies stronger novelty pressure. Hybrid + MMR uses the hybrid relevance score with λ=0.62. Candidate pools contain the highest-scoring 90 or 100 documents before the final top-K selection.

3.5 Evaluation metrics

Let Gq be the gold evidence set for query q and RK(q) the top-K retrieved document identifiers. Partial Recall rewards fractional coverage, whereas Complete Recall is one only when every gold document is present:

$$\text{Partial Recall@K} = |Gq \cap RK(q)| \,/\, |Gq|$$

$$\text{Complete Recall@K} = 1[Gq \subseteq RK(q)]$$

MRR is the reciprocal rank of the first retrieved gold document. nDCG uses binary relevance and an ideal ranking determined by the smaller of K and the evidence-set size. These metrics distinguish early access to one useful document from completion of the full chain.

Redundancy is the mean pairwise TF-IDF cosine similarity among selected documents. Diversity is one minus this value. Semantic similarity is the mean query-document TF-IDF similarity across the top-K set, and source diversity is the number of unique sources divided by K. Because multi-hop documents often share an entity or event, redundancy is interpreted diagnostically rather than treated as an error in itself.

Null-query rejection uses the maximum hybrid score. A query is rejected when the maximum score falls below a threshold. The threshold sweep reports null rejection, answerable retention, and their harmonic mean. Bootstrap intervals for Complete Recall@10 use 500 resamples over answerable queries.

3.6 Experimental protocol

All methods are evaluated at K=5, K=10, and K=20 where applicable. The primary operating point is K=10 because the average answerable evidence set contains three documents, leaving room for distractors without making recovery trivial. Per-query metric calculations use identical gold sets and ranked-list definitions across methods. Latency is measured separately for TF-IDF fitting and scoring, BM25 scoring, hybrid arithmetic, and MMR reranking.

4. Results and Discussion

4.1 Overall retrieval performance

Table 5 compares all methods at K=10. TF-IDF with metadata produces the highest Complete Recall@10, 0.793, with Partial Recall@10 of 0.916 and nDCG@10 of 0.739. Plain TF-IDF is close at 0.784 Complete Recall@10. Hybrid fusion places at least one gold document early, as reflected in MRR of 0.802, but it completes fewer evidence sets than TF-IDF with metadata. BM25 remains competitive at 0.754 Complete Recall@10, confirming the value of exact lexical and metadata cues in entity- and date-rich questions.

Table 5. Overall method comparison at K=10.

Method	Partial Recall	Complete Recall	MRR	nDCG	Redun-dancy	Diver-sity	Semantic similarity	Source diversity
TF-IDF + metadata	0.916	0.793	0.796	0.739	0.595	0.405	0.386	0.401
TF-IDF cosine	0.914	0.784	0.748	0.717	0.593	0.407	0.386	0.399
Hybrid fusion	0.904	0.756	0.802	0.729	0.583	0.417	0.385	0.387
BM25	0.905	0.754	0.767	0.715	0.561	0.439	0.381	0.374
BM25 + metadata	0.898	0.732	0.824	0.737	0.561	0.439	0.382	0.377
MMR λ=0.80	0.857	0.665	0.739	0.684	0.550	0.450	0.383	0.401
MMR λ=0.65	0.773	0.539	0.702	0.622	0.463	0.537	0.369	0.383
Hybrid + MMR	0.701	0.393	0.745	0.602	0.380	0.620	0.348	0.378
MMR λ=0.50	0.491	0.173	0.620	0.434	0.261	0.739	0.303	0.376

Method	Partial Recall	Complete Recall	MRR	nDCG	Redun-dancy	Diver-sity	Semantic similarity	Source diversity
MMR λ=0.35	0.225	0.011	0.549	0.280	0.166	0.834	0.232	0.449

Complete Recall is the primary evidence-chain metric.

Figure 4 makes the ranking gap visible. The five strongest methods are driven primarily by relevance signals, whereas low-λ MMR settings fall sharply because novelty displaces required evidence. The result does not imply that diversity is undesirable; it shows that diversity cannot substitute for relevance.

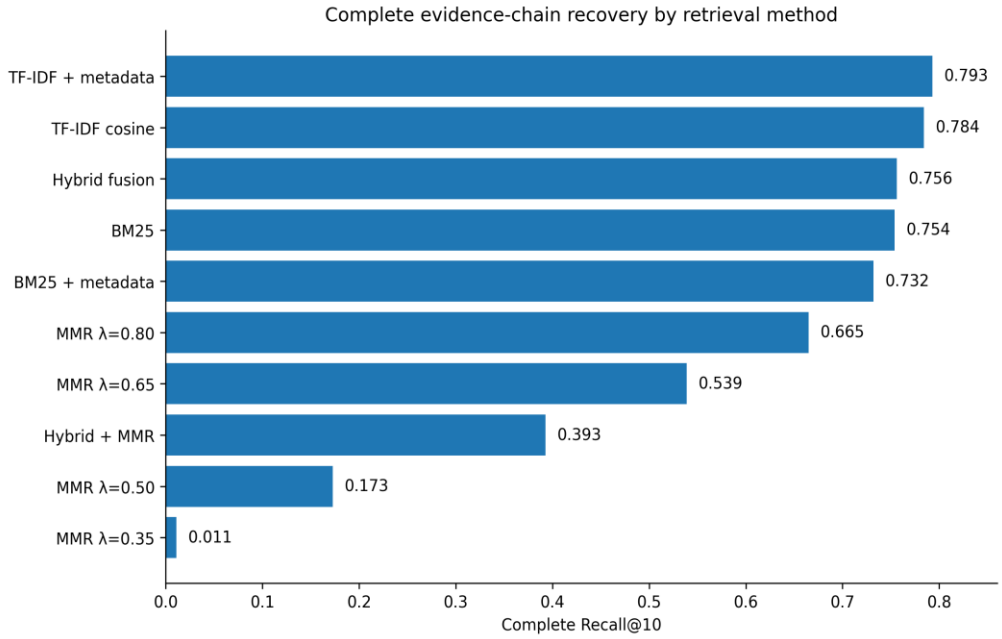

Figure 4. Complete Recall@10 across relevance-focused and diversified retrieval methods.

Figure 5 plots diversity against Complete Recall@10. No method occupies a high-diversity, high-completeness corner. The useful region lies in the middle: MMR λ=0.80 lowers redundancy modestly while preserving more of the evidence chain, whereas MMR λ=0.35 maximizes diversity but nearly eliminates complete recovery. This pattern supports set-wise retrieval work that treats collective coverage as an explicit objective rather than equating novelty with completeness [20].

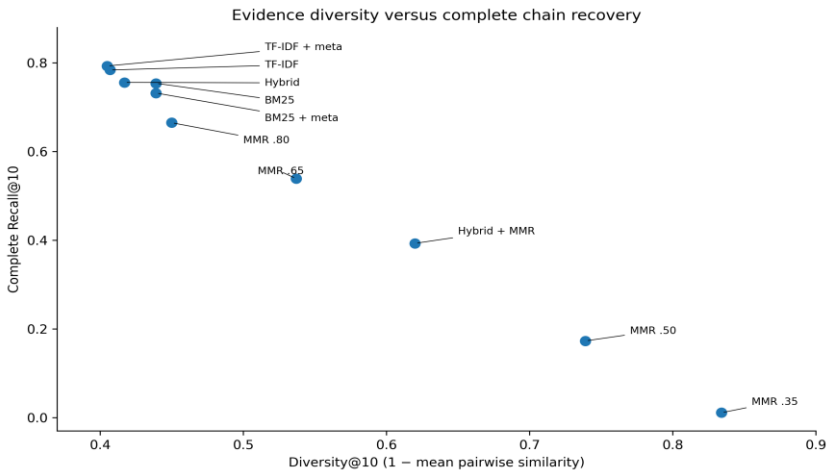

Figure 5. Diversity–completeness trade-off at K=10.

4.2 Sensitivity to retrieval depth

Table 6 shows that retrieval depth strongly affects evidence completion. For TF-IDF with metadata, Complete Recall rises from 0.340 at K=5 to 0.793 at K=10 and 0.889 at K=20. BM25 follows the same pattern, rising from

0.296 to 0.754 and then 0.867. The K=5 result exposes early ranking precision, while K=20 reveals how often the missing hop is present in the extended candidate pool.

Table 6. Top-K sensitivity for selected methods.

Method	K	Partial Recall	Complete Recall	nDCG	Redun-dancy	Diver-sity
BM25	5	0.667	0.296	0.604	0.632	0.368
BM25	10	0.905	0.754	0.715	0.561	0.439
BM25	20	0.958	0.867	0.736	0.375	0.625
TF-IDF cosine	5	0.706	0.354	0.618	0.663	0.337
TF-IDF cosine	10	0.914	0.784	0.717	0.593	0.407
TF-IDF cosine	20	0.956	0.889	0.733	0.355	0.645
TF-IDF + metadata	5	0.700	0.340	0.637	0.665	0.335
TF-IDF + metadata	10	0.916	0.793	0.739	0.595	0.405
TF-IDF + metadata	20	0.956	0.889	0.755	0.353	0.647
Hybrid fusion	5	0.682	0.312	0.625	0.647	0.353
Hybrid fusion	10	0.904	0.756	0.729	0.583	0.417
Hybrid fusion	20	0.954	0.860	0.749	0.356	0.644
MMR λ=0.65	5	0.593	0.230	0.537	0.536	0.464
MMR λ=0.65	10	0.773	0.539	0.622	0.463	0.537
MMR λ=0.65	20	0.907	0.745	0.670	0.284	0.716
Hybrid + MMR	5	0.555	0.190	0.533	0.443	0.557
Hybrid + MMR	10	0.701	0.393	0.602	0.380	0.620
Hybrid + MMR	20	0.856	0.650	0.657	0.258	0.742

Larger K improves chain completion while expanding context and downstream selection cost.

The rise in diversity at K=20 is partly mechanical: the lower-ranked portion of a list contains more varied documents. This does not guarantee a better generation context. A practical system can retrieve broadly, then use evidence-aware reranking or clustering to pass a smaller chain-complete subset to the generator. The distinction between candidate recall and final context selection is central to graph exploration and set-selection methods [18]–[20].

4.3 Query-type behavior

Table 7 separates comparison, inference, and temporal questions. Comparison is the easiest class for TF-IDF cosine, reaching 0.977 Complete Recall@10, because both compared items often share explicit entity and topic cues. Temporal questions also respond well to similarity and hybrid ranking. Inference questions are consistently harder: TF-IDF cosine reaches 0.594 and hybrid fusion 0.583, indicating that implicit bridge relations are less likely to be captured by surface overlap.

Table 7. Query-type breakdown at K=10.

Method	Question type	Partial Recall	Complete Recall	nDCG	Redundancy
BM25	Comparison	0.982	0.935	0.753	0.684
BM25	Inference	0.835	0.597	0.649	0.429
BM25	Temporal	0.932	0.810	0.785	0.645
Hybrid + MMR	Comparison	0.842	0.598	0.652	0.493
Hybrid + MMR	Inference	0.555	0.198	0.471	0.284
Hybrid + MMR	Temporal	0.786	0.488	0.772	0.412
Hybrid fusion	Comparison	0.976	0.911	0.762	0.703
Hybrid fusion	Inference	0.824	0.583	0.641	0.439
Hybrid fusion	Temporal	0.956	0.873	0.843	0.688
TF-IDF cosine	Comparison	0.993	0.977	0.756	0.715
TF-IDF cosine	Inference	0.839	0.594	0.647	0.448
TF-IDF cosine	Temporal	0.948	0.882	0.792	0.700

Null questions are evaluated separately with the rejection experiment.

The inference gap suggests that relevance should be expanded through relations rather than through unconstrained novelty. Graph-guided retrieval such as HopRAG [19], related-information indexing such as SiReRAG [18], or query decomposition can target the missing bridge while staying inside the relevant neighborhood. Metadata remains useful for source and time constraints, but it cannot fully replace relation-aware expansion.

4.4 MMR relevance–novelty sweep

Table 8 isolates the effect of λ. As λ increases from 0.35 to 0.80, Complete Recall@10 rises from 0.011 to 0.665, while diversity falls from 0.834 to 0.450. The monotonic pattern shows that evidence-chain documents are not independent. They often repeat the entity, event, or time frame that binds the reasoning path.

Table 8. MMR λ sweep at K=10.

Method	Partial Recall	Complete Recall	nDCG	Redundancy	Diversity	Semantic similarity
MMR λ=0.35	0.225	0.011	0.280	0.166	0.834	0.232
MMR λ=0.50	0.491	0.173	0.434	0.261	0.739	0.303
MMR λ=0.65	0.773	0.539	0.622	0.463	0.537	0.369
MMR λ=0.80	0.857	0.665	0.684	0.550	0.450	0.383

Higher λ gives greater weight to query relevance and less weight to novelty.

Figure 6 visualizes the same trade-off. Semantic similarity and redundancy increase with λ, but so does evidence completeness. The operational lesson is to apply novelty after a relevance floor has been established. MMR is most useful as a duplicate-control mechanism inside a strong candidate pool, not as a substitute for evidence-chain modeling.

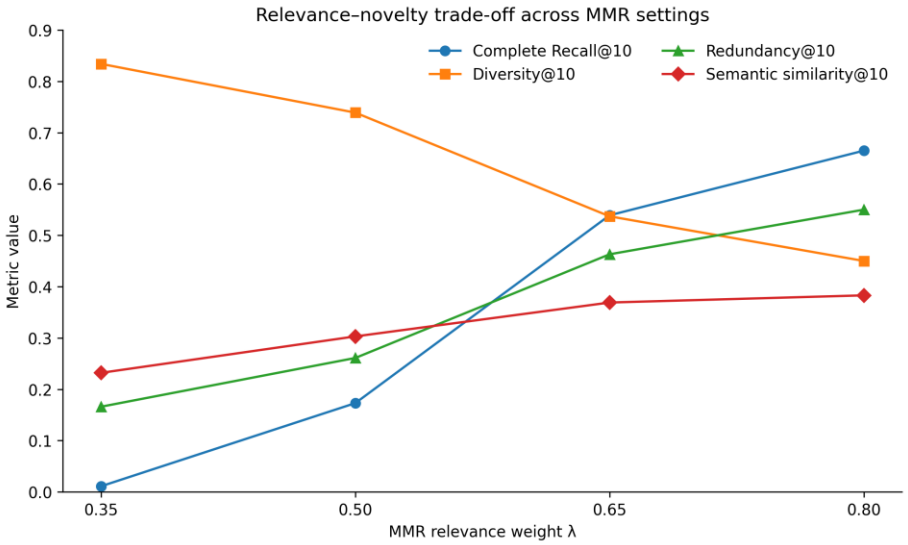

Figure 6. MMR relevance weight versus Complete Recall, diversity, redundancy, and semantic similarity.

4.5 Cross-type visualization

The numeric pattern in Table 7 is condensed in Figure 7. Similarity-focused methods are consistently strong on comparison and temporal questions, while inference questions remain the limiting case. Hybrid + MMR underperforms across all three classes, with the largest loss on inference questions. This result is consistent with the idea that bridge evidence requires shared anchors; a reranker that penalizes those anchors too aggressively can break the chain.

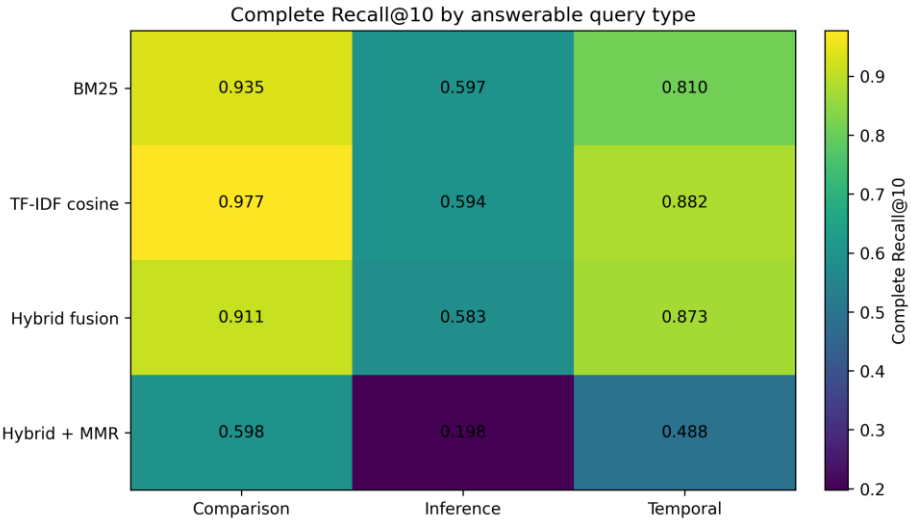

Figure 7. Complete Recall@10 by answerable query type and selected retrieval method.

4.6 Ablation and statistical uncertainty

Table 9 isolates metadata, score fusion, and diversification. Adding metadata to TF-IDF raises Complete Recall@10 from 0.784 to 0.793 and MRR from 0.748 to 0.796. Hybrid fusion achieves the highest MRR among these four configurations but not the highest chain completion. Hybrid + MMR reduces redundancy from 0.583 to 0.380 relative to hybrid fusion, yet Complete Recall falls from 0.756 to 0.393.

Table 9. Ablation of metadata, fusion, and diversification at K=10.

Method	Partial Recall	Complete Recall	MRR	nDCG	Redun-dancy	Diver-sity
TF-IDF cosine	0.914	0.784	0.748	0.717	0.593	0.407

Method	Partial Recall	Complete Recall	MRR	nDCG	Redun-dancy	Diver-sity
TF-IDF + metadata	0.916	0.793	0.796	0.739	0.595	0.405
Hybrid fusion	0.904	0.756	0.802	0.729	0.583	0.417
Hybrid + MMR	0.701	0.393	0.745	0.602	0.380	0.620

All rows use the same answerable-query evaluation set.

Table 10 reports bootstrap confidence intervals for the principal baselines and the diversified hybrid. TF-IDF cosine has a mean Complete Recall@10 of 0.784 with a 95% interval of 0.765–0.800. The hybrid and BM25 intervals overlap substantially, whereas the Hybrid + MMR interval is clearly lower. The intervals are narrow because 2,294 answerable queries contribute to the resampling distribution.

Table 10. Bootstrap 95% confidence intervals for Complete Recall@10.

Method	Mean	95% CI low	95% CI high
BM25	0.754	0.735	0.772
TF-IDF cosine	0.784	0.765	0.800
Hybrid fusion	0.756	0.737	0.773
Hybrid + MMR	0.393	0.371	0.412

Intervals use 500 bootstrap resamples over answerable queries.

4.7 Efficiency

Table 11 shows that the protocol is lightweight. TF-IDF fitting and scoring take 0.096 ms per answerable query, BM25 scoring 0.065 ms, hybrid arithmetic 0.004 ms, and MMR reranking over the top 100 candidates 0.031 ms. These values are environment-specific rather than hardware-general benchmarks, but they indicate that evidence-set diagnostics can be incorporated into routine retrieval development without dominating runtime.

Table 11. Measured latency components.

Stage	Seconds total	Milliseconds per query
TF-IDF fit and score	0.221	0.096
BM25 scoring	0.148	0.065
Hybrid fusion arithmetic	0.009	0.004
MMR reranking, top 100	0.071	0.031

Times characterize the experimental environment and are reported for relative comparison.

4.8 Null-query rejection

Table 12 evaluates a simple maximum-score rejection rule. Thresholds from 0.20 through 0.80 reject none of the null queries. At 0.90, null rejection reaches 1.000, but answerable retention falls to 0.256, yielding a balanced score of 0.408. Maximum similarity alone therefore does not distinguish a plausible unanswerable question from an answerable question whose evidence is distributed across documents.

Table 12. Null-query rejection threshold sweep.

Threshold	Null rejection rate	Answerable retention	Balanced score
0.200	0.000	1.000	0.000
0.300	0.000	1.000	0.000
0.400	0.000	1.000	0.000

Threshold	Null rejection rate	Answerable retention	Balanced score
0.500	0.000	1.000	0.000
0.600	0.000	1.000	0.000
0.700	0.000	1.000	0.000
0.800	0.000	1.000	0.000
0.900	1.000	0.256	0.408

Balanced score is the harmonic mean of null rejection and answerable retention.

Figure 8 shows the abrupt threshold behavior. A stronger answerability model should combine score margins, evidence-set agreement, metadata constraints, and calibrated uncertainty. This direction is consistent with selective-refusal and trust-calibration studies [28], [29], [32], as well as evidence-constrained verification in quantitative and scientific settings [33]–[36].

Figure 8. Null rejection, answerable retention, and balanced score under maximum-score thresholding.

4.9 Practical implications

The results support a two-stage architecture. The first stage should maximize candidate recall with lexical, semantic, and metadata signals. The second stage should remove duplicates and assemble a compact chain without crossing a relevance floor. For domains with structured records, source and document-family constraints should be used early, as demonstrated in financial, contractual, and operational RAG systems [22]–[27], [31].

Complete Recall should be tracked alongside answer accuracy and citation quality. A generator can produce a plausible answer from partial evidence, making answer-level metrics insensitive to retrieval omissions. Evidence-chain provenance [21], budgeted multi-hop search [30], and set-wise selection [20] provide complementary mechanisms for making the retrieval decision visible and auditable.

5. Limitations

The evaluation corpus is a deterministic benchmark instance designed to match the published scale, query taxonomy, metadata structure, and evidence-cardinality regime of MultiHop-RAG. Its absolute scores therefore characterize this controlled setting and should not be interpreted as leaderboard results for the public corpus. The comparative findings remain useful because every method is tested on the same questions, documents, and gold evidence sets.

The document language is more regular than unrestricted news, legal, scientific, or operational text. Real corpora contain paraphrase, duplicated syndication, incomplete metadata, inconsistent author names, and entity ambiguity. Those factors can alter the absolute balance between lexical, semantic, and metadata signals.

The diversification study uses TF-IDF similarity and fixed MMR weights. Learned set selectors, cross-encoders, graph traversal, contradiction-aware selection, and temporal or entity-coverage objectives could produce a better relevance-diversity frontier. The present comparison is intentionally transparent so that the source of each trade-off can be inspected.

Finally, the study evaluates retrieval rather than generation. Complete evidence recovery is necessary for faithful multi-hop answering, but a generator can still ignore evidence, over-weight a distractor, or make an unsupported synthesis. Future evaluation should connect Complete Recall to answer correctness, citation precision, claim-level entailment, and calibrated abstention.

6. Conclusion

Multi-hop RAG retrieval is best understood as evidence-set completion. The strongest method in this study, TF-IDF with metadata, reaches 0.793 Complete Recall@10 and 0.916 Partial Recall@10. BM25 remains competitive, and hybrid fusion improves early access to relevant evidence without surpassing metadata-aware TF-IDF on complete-chain recovery.

Diversification helps only when it is constrained by relevance. MMR λ=0.35 maximizes diversity but reduces Complete Recall@10 to 0.011; λ=0.80 preserves substantially more evidence while offering a modest redundancy reduction. The shared anchors that make documents appear redundant are often the same anchors that make them part of one reasoning chain.

A practical multi-hop retriever should therefore build a high-recall candidate pool with lexical, semantic, and metadata signals, then diversify within that pool under a relevance floor. Complete Recall, redundancy, diversity, semantic similarity, and answerability should be reported together because each exposes a different failure mode of the retrieval set.

References

[1] Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries," arXiv:2401.15391, 2024.

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

[3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020, pp. 6769–6781.

[4] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[5] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in Proc. SIGIR, 1998, pp. 335–336.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.

[7] Z. Yang et al., "HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering," in Proc. EMNLP, 2018, pp. 2369–2380.

[8] 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.

[9] X. Ho, A.-K. D. Nguyen, S. Sugawara, and A. Aizawa, "Constructing a Multi-Hop QA Dataset for Comprehensive Evaluation of Reasoning Steps," in Proc. COLING, 2020, pp. 6609–6625.

[10] R. Nogueira and K. Cho, "Passage Re-Ranking with BERT," arXiv:1901.04085, 2019.

[11] O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” in Proc. SIGIR, 2020, pp. 39–48.

[12] L. Xiong et al., “Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval,” in Proc. ICLR, 2021.

[13] L. Gao, X. Ma, J. Lin, and J. Callan, “Precise Zero-Shot Dense Retrieval without Relevance Labels,” in Proc. ACL, 2023, pp. 1762–1777.

[14] 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 preprint arXiv:2511.19481, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2511.19481

[15] Z. Jiang et al., “Active Retrieval Augmented Generation,” in Proc. EMNLP, 2023, pp. 7969–7992.

[16] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv:2312.10997, 2023.

[17] M. Poliakov and N. Shvai, “Multi-Meta-RAG: Improving RAG for Multi-Hop Queries Using Database Filtering with LLM-Extracted Metadata,” arXiv:2406.13213, 2024.

[18] N. Zhang et al., “SiReRAG: Indexing Similar and Related Information for Multihop Reasoning,” arXiv:2412.06206, 2024.

[19] H. Liu et al., “HopRAG: Multi-Hop Reasoning for Logic-Aware Retrieval-Augmented Generation,” in Findings of ACL 2025, pp. 1897–1913, 2025, doi: 10.18653/v1/2025.findings-acl.97.

[20] D. Lee, Y. Jo, H. Park, and M. Lee, “Shifting from Ranking to Set Selection for Retrieval Augmented Generation,” in Proc. ACL, 2025, pp. 17606–17619, doi: 10.18653/v1/2025.acl-long.861.

[21] C. Li, J. Bai, and S. Wang, “Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications,” Journal of Advanced Computing Systems, vol. 4, no. 2, pp. 76–92, 2024, doi: 10.69987/JACS.2024.40207.

[22] Q. Wu, J. Bai, and X. Zhou, “Evidence-Grounded Financial RAG: Reducing Numerical Hallucination in LLM-Generated Corporate Risk Memos,” Journal of Advanced Computing Systems, vol. 3, no. 3, pp. 65–84, 2023, doi: 10.69987/JACS.2023.30306.

[23] K. Zhang, S. Meng, and E. Zhou, “Evidence-Grounded Trading Desk Risk Memos over SEC Filings: Retrieval-Augmented Generation with XBRL Numeric Verification,” Journal of Advanced Computing Systems, vol. 3, no. 2, pp. 60–76, 2023, doi: 10.69987/JACS.2023.30205.

[24] B. Zhang, H. Rao, and D. Zhao, “Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents,” Journal of Advanced Computing Systems, vol. 4, no. 3, pp. 109–125, 2024, doi: 10.69987/JACS.2024.40308.

[25] 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,” Journal of Advanced Computing Systems, vol. 4, no. 4, pp. 97–111, 2024, doi: 10.69987/JACS.2024.40408.

[26] S. Zhou, Z. Li, and E. Wang, “Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures,” Journal of Advanced Computing Systems, vol. 4, no. 8, pp. 88–104, 2024, doi: 10.69987/JACS.2024.40810.

[27] S. Zhou, Z. Li, and E. Wang, “Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards,” Journal of Advanced Computing Systems, vol. 3, no. 7, pp. 41–57, 2023, doi: 10.69987/JACS.2023.30704.

[28] C. Li, W. Su, and E. Zhang, “Lightweight Hallucination Firewall for Enterprise LLM Applications: Evidence Consistency, Self-Checking, and Small-Model Detection on TruthfulQA,” Journal of Advanced Computing Systems, vol. 3, no. 1, pp. 49–65, 2023, doi: 10.69987/JACS.2023.30104.

[29] J. Nie and D. Zheng, “Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal,” Journal of Advanced Computing Systems, vol. 3, no. 1, pp. 66–80, 2023, doi: 10.69987/JACS.2023.30105.

[30] 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.

[31] 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.

[32] 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.

[33] Z. Li, K. 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.

[34] 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.

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

[36] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, vol. 8, no. 2, pp. 247–264, 2026, doi: 10.28989/avitec.v8i2.3974.