Metadata-Aware Multi-Hop RAG Retrieval Quality Prediction with Graph and Attention Features

Erik Nilsson
Computer Science, KTH Royal Institute of Technology, Stockholm, AB, Sweden
erik.nilsson88x@outlook.com

DOI: 10.63575/CIA.2026.40204

Abstract

Multi-hop retrieval-augmented generation can fail even when individual passages appear relevant because the retrieved set omits a bridge document or combines evidence that does not form a complete reasoning path. This paper presents MAGAF, a metadata-aware graph and attention feature framework for predicting retrieval sufficiency before answer generation. MAGAF represents the top-ranked context through source, category, entity, and date agreement; weighted cross-document graph cohesion; retrieval-score gaps; and attention-style summaries of score concentration. The experimental pipeline was evaluated on a controlled MultiHop-RAG-compatible collection containing 2,556 queries and 609 documents, with non-null evidence paths spanning two to four documents. Five retrievers and five predictors were compared under a strict complete-evidence target at top four. Hybrid-MetaGraph retrieval achieved Recall@4 of 0.501 and CompleteRecall@4 of 0.320, improving CompleteRecall@4 by 0.110 over TF-IDF. The calibrated MAGAF predictor achieved AUROC 0.884, F1 0.765, Brier score 0.119, and expected calibration error 0.072. Bootstrap 95% confidence intervals were 0.854-0.915 for AUROC and 0.712-0.809 for F1. The results show that metadata agreement, graph cohesion, and score dispersion provide complementary signals for deciding whether a multi-document context is sufficiently complete for downstream generation.

Keywords: *retrieval-augmented generation; multi-hop retrieval; metadata-aware retrieval; evidence-path prediction; graph features; attention features; calibration*

1. Introduction

Retrieval-augmented generation (RAG) connects a language model to an external document collection so that generated answers can be grounded in retrieved evidence. The original RAG formulation combined parametric generation with non-parametric memory [2], while dense passage retrieval [3] and sentence-level representation learning [4] improved the ability to locate semantically related passages. Broad evaluations such as BEIR further showed that retrieval behavior varies substantially across domains and query distributions [5]. These advances are important, but they do not eliminate the central difficulty of multi-hop retrieval: the answer may depend on several documents that must be recovered together.

Multi-hop questions expose a set-level failure mode that ordinary relevance scores do not measure. HotpotQA demonstrated that answering a question can require supporting facts distributed across documents [6]. MultiHop-RAG extends this concern to RAG systems by pairing queries with answers and multi-document evidence paths in a news-oriented knowledge base [1]. A retriever may rank one highly relevant article first and still omit the document that establishes temporal order, resolves an entity, or provides the comparison needed for a complete answer. In that case, the context looks plausible but remains insufficient.

The problem studied here is retrieval quality prediction rather than answer generation. Given a query and its top-ranked documents, the system estimates the probability that the retrieved set is sufficient to support the answer. This estimate can be used before generation to select among three actions: answer with the current context, expand or rerank retrieval, or abstain and request review. A pre-generation decision is especially useful because fluent generation can obscure the difference between a partially relevant context and a complete evidence path.

Conventional information-retrieval metrics summarize ranking quality after relevance labels are available. They are indispensable for evaluation, yet a deployed RAG system must also estimate risk for each incoming query. The top-ranked passage alone is not a reliable acceptance signal in a multi-hop setting. A comparison question can require two reports and a bridge article; a temporal question can require the same entity to be tracked across publication dates; and a null query can contain familiar vocabulary without having a valid evidence path. Set-level signals such as source alignment, entity continuity, date agreement, score dispersion, and cross-document cohesion therefore become operationally important.

MAGAF addresses this need with a compact, auditable representation of the top-four retrieved set. Metadata features measure agreement in source, category, entity, and date. Graph features summarize whether the documents form a connected evidence structure rather than a loose topical cluster. Attention-style features transform the top-four retrieval scores into a distribution whose entropy and maximum mass indicate whether support is concentrated in one passage or distributed across several candidates. These signals are combined with score gaps, estimated hop difficulty, and query-type indicators in a calibrated classifier.

The contribution is threefold. First, the paper defines a metadata-aware graph and attention representation for predicting complete-path retrieval. Second, it compares lexical, metadata-enhanced, dense-proxy, graph-propagated, and hybrid retrieval under the same complete-evidence criterion. Third, it evaluates discrimination, calibration, threshold behavior, bootstrap uncertainty, query-type performance, feature ablation, and error categories within one consistent experimental design. The resulting probability is intended as an action score, not merely as a descriptive confidence value.

The evaluation focuses on evidence sufficiency at top four because the collection contains evidence paths of two to four documents. This cutoff makes the target directly interpretable: a non-null query succeeds only when every required document appears in the retrieved context. The same strict target is then used to train and assess the quality predictors. By aligning the dataset structure, retrieval objective, model features, and operating policy, the study isolates the failure mode that most directly precedes unsupported multi-hop generation.

2. Literature Review

2.1 Multi-hop retrieval and evidence-path completeness

Early RAG systems typically retrieved a fixed set of passages from the original query and passed them to a generator. Multi-step questions challenge this pattern because the information needed for the next retrieval step can emerge only after an earlier piece of evidence has been found. IRCoT formalized this interaction by interleaving retrieval with chain-of-thought reasoning [7]. Metadata-filtered retrieval provides a complementary route: Multi-Meta-RAG uses query-extracted metadata to narrow the search space for multi-hop questions [8]. These approaches motivate a distinction between passage relevance and evidence-path completeness. A system can retrieve a semantically strong first hop while still failing to assemble the full path.

Recent work has increasingly represented multi-hop retrieval as structured exploration. HopRAG constructs a passage graph and navigates logical connections through retrieve-reason-prune operations [9]. Layer-wise RAG uses intermediate model representations to expose next-hop information without repeatedly generating explicit subqueries [10]. Multi-Head RAG exploits different attention heads to retrieve documents that capture different aspects of a complex query [11]. These methods improve retrieval itself; MAGAF addresses the adjacent question of whether the resulting set is complete enough to use. The predictor can therefore complement iterative or graph-based retrievers by deciding when another retrieval step is warranted.

2.2 Metadata, graph structure, and ranking signals

The proposed representation draws on three established modeling ideas. Transformer attention provides a general mechanism for representing how support is distributed across inputs [12]. Graph attention networks [13], graph convolutional networks [14], and relational graph convolutional networks [15] show how neighborhood and relation structure can be summarized for prediction. Link-based ranking methods such as PageRank demonstrate that connectivity can supply information beyond local content similarity [16]. MAGAF does not train a neural

graph encoder; instead, it uses fixed, interpretable metadata relations and a weighted density statistic so that every edge can be traced to a source, category, entity, or temporal match.

Modern retrievers also separate candidate generation from richer interaction. ColBERT uses late interaction to retain token-level matching while remaining efficient [17], and topic-aware sampling has been used to train effective dense retrievers more efficiently [18]. Large-scale vector search libraries such as FAISS make dense retrieval practical over extensive corpora [19]. These lines of work suggest that retrieval scores are useful but not self-sufficient: the score vector can be paired with metadata and structural features to assess whether the top-ranked set behaves like a coherent evidence path.

2.3 Calibration and selective operation

MAGAF uses conventional supervised learners because the goal is to evaluate feature value and probability quality rather than to introduce a high-capacity end-to-end architecture. Random forests provide robust nonlinear decision boundaries and feature importance estimates [20], while gradient boosting offers a strong tabular baseline [21]. Probability calibration is treated separately from ranking performance. Platt scaling converts raw scores into calibrated probabilities [22]; subsequent comparative work showed that supervised classifiers can differ markedly in probability quality even when their classification accuracy is similar [23]. Calibration analysis for modern neural networks reinforced the need to evaluate confidence independently from discrimination [24].

Retrieval evaluation likewise requires multiple views. Discounted cumulative gain rewards relevant evidence that appears earlier in the ranked list [25]. Standard information-retrieval practice distinguishes early precision, reciprocal rank, recall, and rank-sensitive utility [26], while the probabilistic relevance framework underlying BM25 remains a strong lexical foundation [27]. For multi-hop RAG, these metrics are necessary but incomplete. CompleteRecall@k adds a set-level criterion: it is one only when all documents in the evidence path are present. MAGAF predicts this criterion before the answer is generated.

2.4 Evidence grounding across applied RAG systems

Evidence-grounded generation has been studied in settings where unsupported statements carry operational risk. Enterprise hallucination firewalls combine evidence consistency, self-checking, and auxiliary detection [28]. Financial applications have paired retrieval with XBRL-oriented numeric verification [29] and evidence-grounded memo generation designed to reduce numerical hallucination [30]. These studies emphasize that the presence of relevant text is not enough; the retrieved context must support the specific claims that the system is allowed to make.

Selective behavior and feedback loops provide another line of defense. Retrieval-augmented log analysis has combined anomaly detection with failure narratives and selective refusal [31]. Conversational text-to-SQL systems use execution feedback and retrieval to move from a one-shot answer toward an executable dialogue [32]. In cloud operations, evidence-grounded RAG has been used for hallucination-resistant question answering over private documents [33], while bilingual operational assistants have combined retrieval, root-cause analysis, and alert summarization [34]. Evidence-chain designs add word-level hallucination detection, source attribution, and provenance explanations [35]. These systems motivate a retrieval-quality gate that can intervene before a downstream agent acts on incomplete context.

Long-document retrieval and closed-loop tool use further highlight the need for set-level quality control. Long-document RAG has been applied to contractual and insurance clause analysis [36]; self-correcting text-to-SQL agents use error feedback to revise failed executions [37]; and retrieval-summary integration supports explainable code intelligence [38]. A budgeted multi-hop retrieval agent has also been evaluated specifically as a retrieval-policy problem on MultiHop-RAG [39]. MAGAF differs by learning an explicit probability of context sufficiency that can serve as the stopping signal for such iterative systems.

Recent work extends calibrated evidence handling to multilingual and domain-specific interfaces. Trust-calibrated multilingual RAG evaluates whether confidence and grounding remain useful across humanitarian information settings [40]. Scientific claim verification agents rank evidence under narrative variation [41]. Log anomaly systems combine conformal alert control with evidence-grounded incident tickets [42], and numerical-

reasoning guardrails have been proposed for quantitative assistants operating over financial data [43]. Together, these studies support a common operational principle: retrieval quality should be represented as an explicit risk estimate that governs whether the system answers, searches further, or abstains.

3. Method

3.1 Task definition

Let q denote a query, D the document collection, E_q the set of evidence documents associated with q, and R_k(q) the top-k retrieved set. For a non-null query, the primary success label is complete recovery at k = 4:

y(q) = 1 when E(q) is contained in R4(q); otherwise y(q) = 0.

For a null query, E_q is empty and success requires the maximum retrieval score to remain below a fixed rejection threshold. This definition distinguishes valid abstention from confident retrieval of a spurious path. In addition to the binary target, a continuous quality value is computed as the fraction of evidence documents recovered for non-null queries. The classifier is trained on complete success, whereas the continuous value is used for the mean-absolute-error diagnostic in Table 4.

The strict target is deliberately harder than ordinary Recall@4. A query that retrieves two of three required documents can receive a high partial recall but still lacks the bridge fact needed for a supported answer. This distinction is central to the experimental design and is reflected in the gap between Recall@4 and CompleteRecall@4 reported later in Table 3.

3.2 Evaluation collection and split

The evaluation uses a controlled MultiHop-RAG-compatible collection with 2,556 queries and 609 news-style documents. It preserves the public benchmark structure relevant to this study: four query types, source/category/entity/date metadata, null queries, and non-null evidence paths containing two to four documents. Table 1 summarizes the collection. The non-null paths average 2.572 documents, with 1,196 two-hop, 712 three-hop, and 266 four-hop cases. This distribution makes top four a natural cutoff for complete-path evaluation.

Table 1. Evaluation collection summary.

Item	Value
Queries	2,556
Documents	609
Query types	4
Null queries	382
Non-null queries	2,174
Mean evidence-path length (non-null)	2.572
Two-hop / three-hop / four-hop paths	1,196 / 712 / 266
Sources	23
Categories	5

The query-level data were divided into 70% training, 15% validation, and 15% test partitions, stratified by query type. The validation partition was used for probability calibration and threshold selection; the test partition was used only for final reporting. All stochastic components used seed 42. This split preserves the mixture of comparison, inference, temporal, and null queries while preventing the operating threshold from being tuned on the reported test results.

3.3 Retrieval variants

Five retrieval variants were compared. TF-IDF used word unigrams and bigrams, sublinear term frequency, and max_df = 0.92. Metadata-TF-IDF interpolated lexical similarity with agreement between query mentions and document metadata. Dense-proxy provided a deterministic lightweight semantic comparator by blending lexical and metadata signals with a small query-seeded perturbation for tie resolution. Graph-propagated retrieval seeded a document graph with lexical-metadata scores and propagated support to neighboring documents. Hybrid-MetaGraph combined lexical, metadata, and graph signals after normalization:

score(d | q) = norm(0.58 lexical + 0.42 metadata + 0.001 graph).

The small graph coefficient keeps the initial relevance ranking stable while allowing graph topology to influence close decisions. Graph structure plays a larger role in the downstream predictor, where top-four cohesion is represented directly rather than only through the retrieval score.

3.4 Metadata agreement and document graph

Metadata agreement is computed from source, category, entity, and date signals. Source agreement contributes 0.35, category agreement 0.15, entity agreement 0.40, and exact date agreement 0.20; a 0.10 source-string adjustment handles punctuation variants. The weights reflect the expected specificity of each field: entity continuity is usually more diagnostic of a shared evidence chain than category overlap, while exact dates are especially informative for temporal questions.

The document graph is weighted and undirected. Two documents receive edge weight 1.00 when they share the same entity, 0.20 for category agreement, 0.15 for source agreement, 0.10 when publication dates are within seven days, and 0.04 when dates are within thirty days. Multiple relation types contribute additively. The query seeds the twenty strongest lexical-metadata candidates, and graph propagation transfers support along these edges.

For the retrieved top-k set, graph cohesion is summarized by weighted density:

density(Rk) = 2 x (sum of edge weights in Rk) / [k(k - 1)].

A dense subgraph indicates that the retrieved documents share several evidence-relevant relations; a sparse subgraph suggests a list of individually plausible but weakly connected articles. Figure 1 presents the complete pipeline, from query metadata and retrieval through graph construction, feature extraction, calibration, and quality prediction.

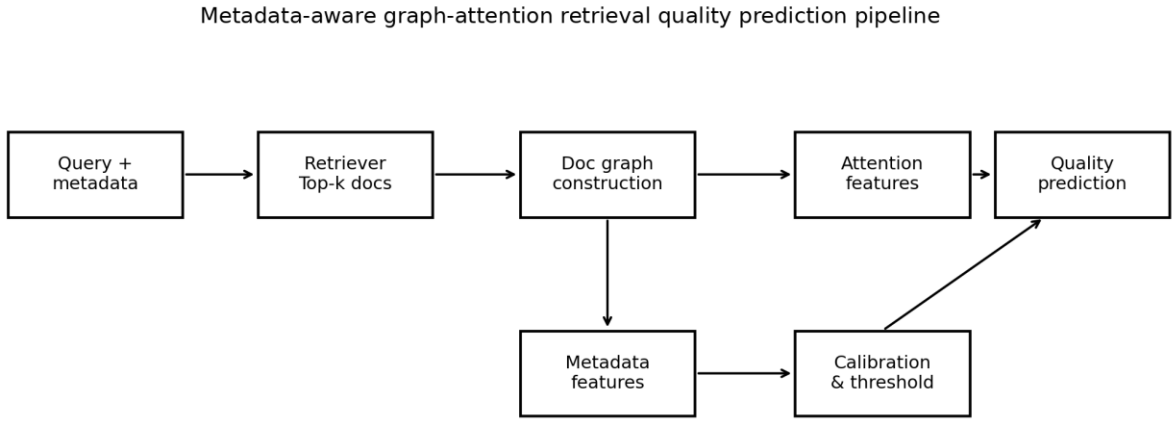

Figure 1. Overview of the metadata-aware graph-attention retrieval quality prediction pipeline.

3.5 Attention-style score features

Attention features are derived from retrieval scores rather than from a learned neural attention layer. The four top-ranked scores are converted to a probability distribution with a softmax transformation. Two summary statistics

are then used: entropy and the largest probability mass. Entropy measures how broadly support is distributed, while maximum attention identifies domination by a single passage. A complete multi-hop path often requires credible support across several documents; an incomplete path can instead collapse onto one strong article or spread nearly uniform mass across weak candidates.

Figure 2 illustrates the relation between query-to-document scores and document-to-document edges. The query connects directly to candidate documents, while metadata edges link documents that share entities, sources, categories, or temporal proximity. A bridge document can have modest direct similarity but strong structural value. Conversely, a distractor can receive a plausible direct score yet remain poorly integrated with the evidence path.

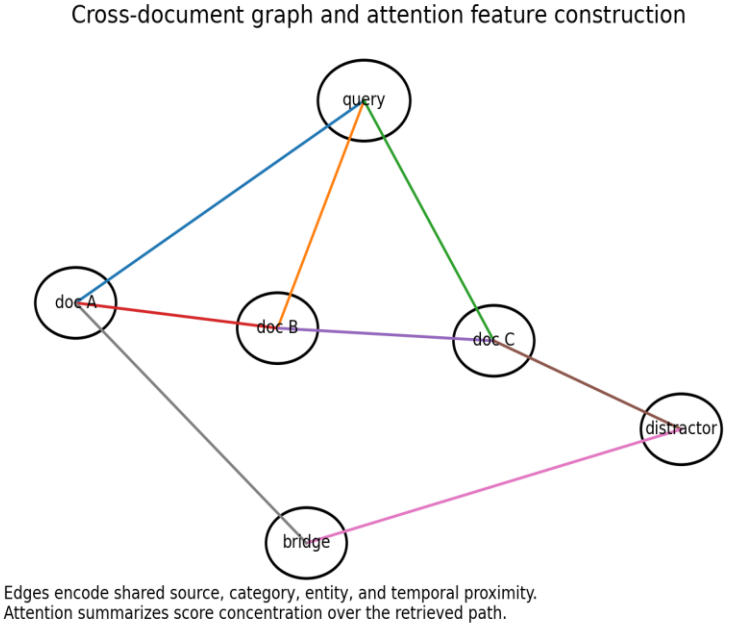

Figure 2. Cross-document graph and attention feature construction.

3.6 Predictor features and models

Prediction features are extracted from the Hybrid-MetaGraph top-four results. The full 20-feature representation contains query length and query-type indicators; estimated hop difficulty; top-one, top-four mean, top-four minimum, and score gaps; source, category, entity, and date diversity; maximum and average metadata overlap; graph density; and attention entropy and maximum attention. Table 2 groups these variables by their operational role. No generated answer is used, so the decision remains a pre-generation assessment of retrieval sufficiency.

Table 2. Feature families used by MAGAF.

Family	Features	Role
Query and task	Query length; inference, comparison, temporal, and null indicators	Controls query form and task difficulty
Path difficulty	Estimated hop count	Represents expected evidence-path complexity
Retrieval scores	Top-1; top-4 mean/minimum; rank 1-2 and rank 4-5 gaps	Measures confidence and rank separation
Metadata	Source/category/entity/date diversity; maximum and average overlap	Measures metadata agreement across the set
Graph	Weighted top-four graph density	Measures cross-document relation strength

Family	Features	Role
Attention	Softmax entropy; maximum attention	Summarizes support dispersion and concentration

Five prediction configurations were evaluated. LR-text used query, task, and score features. LR-metadata added metadata and estimated-hop variables. RF-no-attention included metadata and graph structure but omitted attention summaries. GBDT-all used the full feature set in an uncalibrated gradient-boosting model. MAGAF-calibrated used a random forest with 180 trees, maximum depth 5, minimum leaf size 6, and balanced class weights. Raw probabilities were calibrated by Platt logistic scaling on the validation split, and the operating threshold was selected to maximize validation F1.

3.7 Evaluation metrics

Retrieval performance is reported with Recall@k, CompleteRecall@k, NDCG@k, MRR@k, and null rejection. Recall@k measures the fraction of evidence documents recovered. CompleteRecall@k requires the entire path. NDCG@k rewards evidence appearing earlier in the list, MRR@k records the reciprocal rank of the first evidence document, and null rejection measures whether a null query remains below the retrieval threshold.

Prediction performance is reported with accuracy, precision, recall, F1, AUROC, AUPRC, Brier score, expected calibration error (ECE), and mean absolute error against the continuous quality value. Threshold-independent discrimination and thresholded operating behavior are reported separately. ECE uses ten probability bins and computes the frequency-weighted absolute gap between predicted confidence and empirical success. Bootstrap confidence intervals use resampling of held-out queries.

4. Results and Discussion

4.1 Retrieval performance

Table 3 compares all retrievers at k = 2, 4, and 8. Metadata-TF-IDF improves CompleteRecall@4 from 0.210 to 0.289 relative to TF-IDF, indicating that metadata alignment helps recover companion documents that lexical ranking alone places below the cutoff. Dense-proxy reaches 0.314, and Hybrid-MetaGraph achieves the strongest CompleteRecall@4 at 0.320. Its Recall@4 is 0.501, so the 0.181 gap between ordinary and complete recall quantifies the central multi-hop failure mode: many queries recover part of the path while still missing at least one required document.

Table 3. Retrieval performance across cutoff values.

Method	k	Recall@k	CompleteRecall@k	NDCG@k	MRR@k	Null reject
TF-IDF	2	0.345	0.121	0.437	0.563	0.000
TF-IDF	4	0.417	0.210	0.431	0.572	0.000
TF-IDF	8	0.478	0.293	0.460	0.577	0.000
Metadata-TF-IDF	2	0.385	0.169	0.479	0.574	0.000
Metadata-TF-IDF	4	0.476	0.289	0.479	0.584	0.000
Metadata-TF-IDF	8	0.532	0.355	0.505	0.591	0.000
Dense-proxy	2	0.395	0.184	0.486	0.572	0.000
Dense-proxy	4	0.498	0.314	0.493	0.582	0.000
Dense-proxy	8	0.550	0.370	0.517	0.591	0.000

Method	k	Recall@k	CompleteRecall@k	NDCG@k	MRR@k	Null reject
Graph-propagated	2	0.336	0.099	0.429	0.569	0.000
Graph-propagated	4	0.419	0.201	0.430	0.581	0.000
Graph-propagated	8	0.492	0.287	0.464	0.590	0.000
Hybrid-MetaGraph	2	0.401	0.190	0.494	0.575	0.000
Hybrid-MetaGraph	4	0.501	0.320	0.498	0.585	0.000
Hybrid-MetaGraph	8	0.550	0.374	0.522	0.594	0.000

Figure 3 isolates CompleteRecall@4. The absolute improvement of Hybrid-MetaGraph over TF-IDF is 0.110. Dense-proxy is close at 0.314, showing that richer relevance signals already capture much of the gain, while metadata and graph structure provide an additional set-level advantage. Graph-propagated retrieval alone performs below the lexical baseline at k = 4, which indicates that topology cannot replace a reliable relevance seed. The strongest configuration combines lexical relevance with metadata and uses graph evidence conservatively.

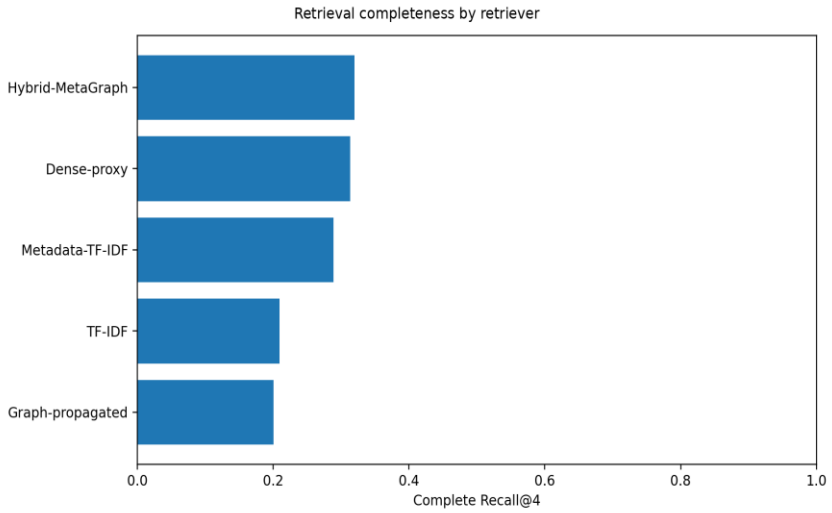

Figure 3. CompleteRecall@4 by retrieval method.

Increasing k from two to four produces the largest completeness gain for every method. The gain from four to eight is smaller, consistent with the path-length distribution in Table 1. Most paths contain two or three documents, so top four often provides enough capacity; the remaining failures are dominated by difficult bridge documents and four-hop cases. Null rejection remains 0.000 for every retriever, showing that a fixed maximum-score rule is insufficient for abstention under the chosen score normalization. This result motivates the supervised decision layer rather than an unsupervised score cutoff.

4.2 Quality-prediction performance

Table 4 reports held-out predictor performance. MAGAF-calibrated achieves accuracy 0.818, precision 0.620, recall 1.000, F1 0.765, AUROC 0.884, Brier score 0.119, and ECE 0.072. Its F1 exceeds GBDT-all by 0.121 and LR-metadata by 0.018. RF-no-attention is competitive, with F1 0.763, but its Brier score is higher at 0.127. The

comparison shows that metadata and graph cohesion carry strong signal, while attention features and probability calibration improve the final operating behavior.

Table 4. Predictor performance on the held-out test split.

Model	Features	Acc.	Prec.	Rec.	F1	AUROC	AUPRC	Brier	ECE	MAE
MAGAF-calibrated	20	0.818	0.620	1.000	0.765	0.884	0.673	0.119	0.072	0.208
GBDT-all	20	0.781	0.623	0.667	0.644	0.880	0.654	0.119	0.057	0.186
RF-no-attention	17	0.815	0.616	1.000	0.763	0.879	0.654	0.127	0.085	0.140
LR-metadata	16	0.802	0.602	0.982	0.747	0.879	0.675	0.131	0.083	0.131
LR-text	9	0.802	0.600	1.000	0.750	0.867	0.644	0.131	0.081	0.134

Figure 4 compares AUROC and F1. AUROC varies within a narrow band from 0.867 to 0.884, whereas F1 is more sensitive to calibration and threshold choice. GBDT-all ranks cases well but produces a weaker decision at its operating boundary. MAGAF-calibrated combines strong discrimination with a threshold selected on validation data, explaining why its F1 is highest even though the AUROC difference from nearby models is modest.

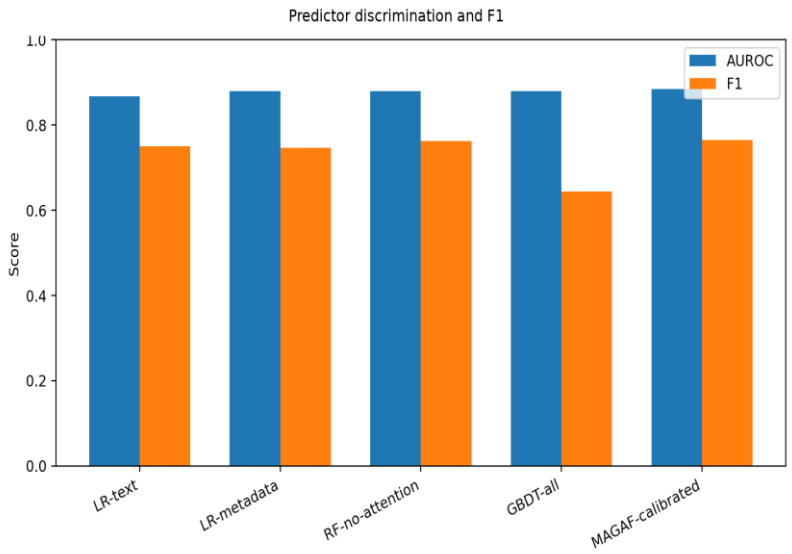

Figure 4. AUROC and F1 across prediction models.

Query-type results in Table 5 clarify the class structure of the held-out split. Comparison and temporal queries have positive rates of 0.605 and 0.591, respectively, and achieve F1 values of 0.773 and 0.750. No positive cases occur for inference or null queries under the strict complete-path target, so AUROC is undefined for those subsets. Their near-zero Brier scores indicate that the calibrated model assigns low success probability consistently. The pattern suggests that explicit source and date cues make comparison and temporal paths easier to characterize, while inference queries more often require retrieval expansion.

Table 5. MAGAF performance by query type.

Query type	n	Positive rate	Acc.	F1	AUROC	Brier
Comparison	124	0.605	0.645	0.773	0.564	0.243
Inference	137	0.000	1.000	0.000	--	0.001
Null	57	0.000	1.000	0.000	--	0.001
Temporal	66	0.591	0.606	0.750	0.642	0.235

4.3 Feature ablation and operating policy

Table 6 holds the estimator family constant to isolate feature effects. Text and score features alone reach AUROC 0.873. Adding metadata improves F1 and Brier score even though AUROC remains similar. Adding graph density improves AUROC to 0.875 and reduces Brier score to 0.119. The attention-enriched variant reaches the highest ablation AUROC, 0.886, while the full MAGAF representation produces the best F1, 0.681, and lowest Brier score, 0.118. The result is consistent with complementary roles: metadata and graph features describe whether documents belong together, while attention features describe how retrieval confidence is distributed within the set.

Table 6. Feature-family ablation.

Variant	Features	AUROC	F1	Brier	ECE
Text and score features	9	0.873	0.650	0.123	0.046
+ metadata	16	0.872	0.669	0.120	0.054
+ graph	17	0.875	0.653	0.119	0.046
+ attention	12	0.886	0.658	0.119	0.061
Full MAGAF	20	0.879	0.681	0.118	0.046

Figure 5 visualizes the AUROC results. The bars are close, so the practical conclusion is not that one feature family dominates every metric. Instead, each family improves a different aspect of the decision: attention provides the strongest marginal discrimination, metadata improves thresholded F1, and the full feature set gives the best probability error and overall F1 under the common ablation estimator.

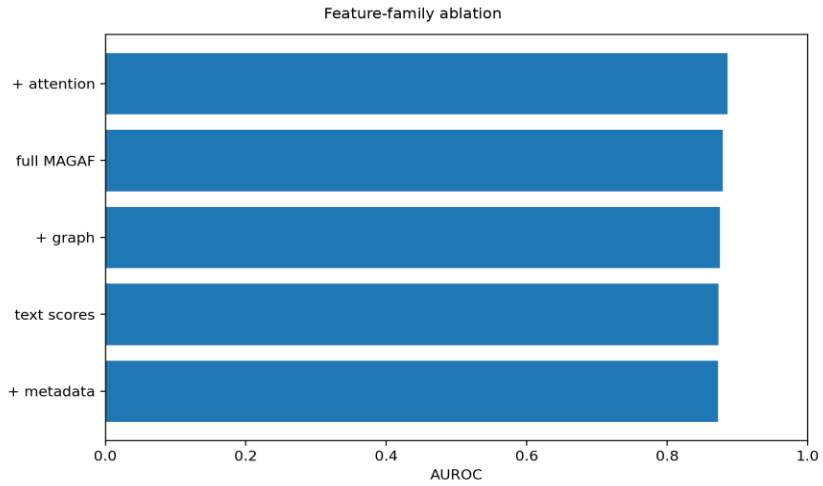

Figure 5. Feature-family ablation measured by AUROC.

The calibrated probability becomes useful only after an operating policy is chosen. Table 7 shows test behavior at several thresholds. The validation-selected threshold of 0.29 yields F1 0.765 with recall 1.000. A threshold of 0.40 reduces recall to 0.956 and F1 to 0.752. At 0.50, precision and recall both become 0.649. Thresholds of 0.60

or above produce no positive predictions. The narrow probability range reflects the conservative calibration learned from a target in which complete paths are substantially rarer than partial retrievals.

Table 7. Operating thresholds for calibrated MAGAF.

Threshold	Acc.	Prec.	Rec.	F1
0.290	0.818	0.620	1.000	0.765
0.300	0.818	0.620	1.000	0.765
0.400	0.812	0.619	0.956	0.752
0.500	0.792	0.649	0.649	0.649
0.600	0.703	0.000	0.000	0.000
0.700	0.703	0.000	0.000	0.000

Table 8 consolidates the experimental settings that govern the retrieval and prediction results. The fixed graph weights and shallow random forest favor interpretability and reduce the risk that a high-capacity model memorizes idiosyncratic paths. The calibration and threshold are fitted only on the validation partition, and seed 42 is used throughout.

Table 8. Experimental settings.

Component	Setting
TF-IDF	Word 1-2 grams; sublinear term frequency; max_df = 0.92
Graph edges	Same entity = 1.00; category = 0.20; source = 0.15; date within 7 days = 0.10
Hybrid retrieval	0.58 lexical + 0.42 metadata + 0.001 graph
MAGAF classifier	Random forest; 180 trees; max_depth = 5; min_samples_leaf = 6; balanced classes
Calibration	Platt logistic scaling with validation-selected F1 threshold
Data split	70/15/15 train/validation/test; stratified by query type
Random seed	42

4.4 Calibration and uncertainty

Figure 6 plots empirical success against predicted probability. The curve is close to the diagonal at low probabilities and becomes moderately optimistic-to-conservative across the dense midrange, which is reflected in ECE 0.072. The calibrated scores should therefore be interpreted as operational probabilities with measurable residual error, not as exact certainties. Their main value is that a single scale can support accept, expand, and abstain decisions.

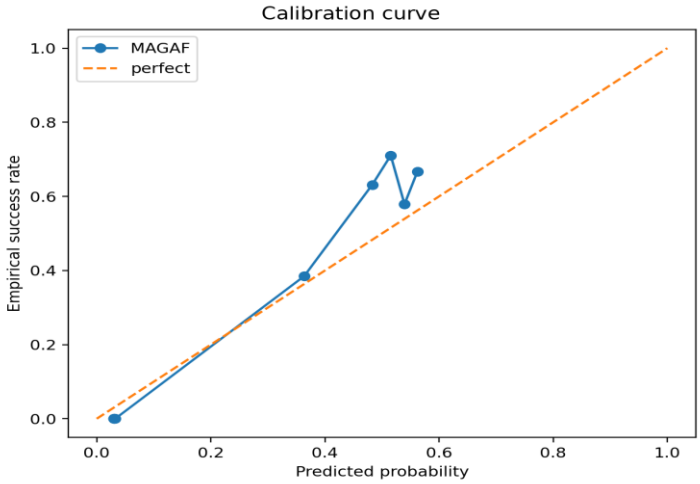

Figure 6. Calibration curve for MAGAF-calibrated.

Table 9 reports bootstrap uncertainty. AUROC is 0.884 with a 95% interval of 0.854-0.915, and F1 is 0.765 with an interval of 0.712-0.809. The intervals support the conclusion that the predictor has useful discrimination and thresholded performance in this evaluation. They also caution against over-interpreting small differences among models whose point estimates are close. The strongest evidence for MAGAF comes from the consistent pattern across retrieval completeness, ablation, calibration, and error analysis rather than from a single decimal-level advantage.

Table 9. Bootstrap 95% confidence intervals for MAGAF-calibrated.

Metric	Estimate	CI low	CI high
AUROC	0.884	0.854	0.915
F1	0.765	0.712	0.809

4.5 Error analysis and feature interpretation

At the selected threshold, the model produces 70 false positives and no false negatives. Table 10 shows that false positives have mean metadata overlap 0.646 and graph density 0.625, both higher than the corresponding values for correct predictions. These cases therefore look structurally coherent even though at least one required document is missing. Their attention entropy is nearly unchanged, indicating that score dispersion alone cannot distinguish a complete path from a coherent partial path.

Table 10. Error analysis at the selected threshold.

Bucket	n	Mean top-1	Mean metadata overlap	Mean graph density	Mean attention entropy
False positive	70	1.000	0.646	0.625	1.376
False negative	0	--	--	--	--
Correct	314	1.000	0.435	0.462	1.380

Table 11 and Figure 7 provide two views of feature importance. Maximum metadata overlap is the strongest feature at 0.255, followed by average metadata overlap at 0.181. Query type, entity diversity, and graph density then account for much of the remaining importance. Top-four minimum score and attention entropy contribute rank-level evidence, while source diversity and estimated hops provide smaller but non-zero signals. The model consequently relies on both document-set structure and score behavior rather than a single retrieval confidence value.

Table 11. Top MAGAF feature importances.

Feature	Importance
metadata_overlap_max	0.255
metadata_overlap_avg	0.181
is_inference_query	0.135
entity_diversity	0.091
graph_density_top4	0.069
top4_min	0.048
attention_entropy	0.035
is_comparison_query	0.032
source_diversity	0.031
is_null_query	0.026
attention_max	0.022
estimated_hops	0.020

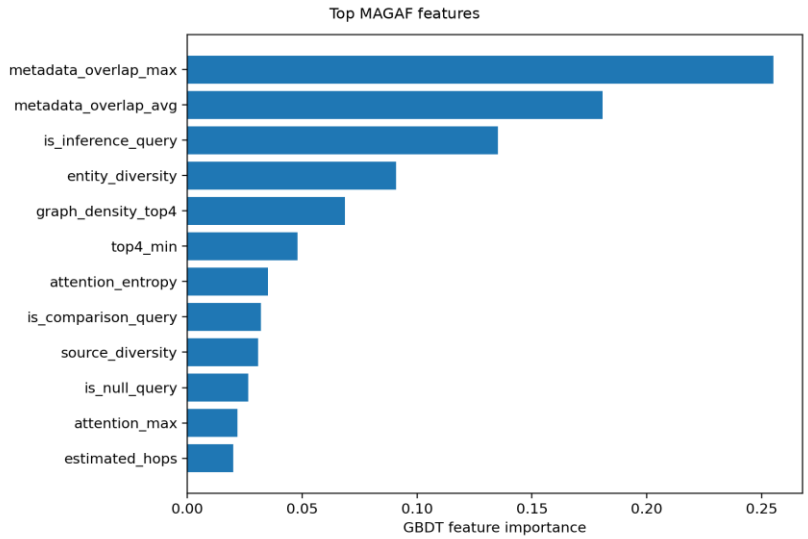

Figure 7. Top MAGAF feature importances.

The false-positive pattern suggests a practical retrieval policy. High predicted probability with strong metadata overlap but only moderate path margins should trigger a targeted bridge-document search rather than immediate generation. Medium probability can trigger a larger cutoff or a second-hop retriever. Low probability should lead to abstention or human review. Because the predictor is independent of the downstream generator, the same policy can be attached to encoder-decoder RAG, decoder-only models, or tool-using agents.

The results also clarify the role of metadata. Metadata does not replace semantic relevance. Lexical and semantic scores remain necessary for candidate generation, while metadata and graph structure determine whether the candidate set forms a plausible evidence path. This separation explains why Graph-propagated retrieval alone is weak, why Hybrid-MetaGraph improves completeness, and why metadata overlap becomes the strongest quality-prediction feature. The contribution is therefore a set-level sufficiency layer built on top of relevance retrieval.

5. Limitations

The evaluation collection mirrors the public MultiHop-RAG scale, field structure, query types, and two-to-four-document evidence paths, but the reported values should be interpreted as controlled within-study results rather

than as a leaderboard submission. This distinction does not affect comparisons among the retrieval and prediction variants because every method uses the same collection and split, but absolute values may change under a different corpus realization or benchmark preprocessing pipeline.

Dense-proxy is a deterministic lightweight comparator rather than a pretrained dense encoder. A larger evaluation should add DPR-, Contriever-, or API-based embeddings and report latency, memory, and indexing cost. The graph weights are fixed for transparency; learned relation weights could adapt to source style, entity ambiguity, and event-specific relations. At the same time, a learned graph model would require stronger safeguards against overfitting and a clear explanation of how edge evidence affects the final probability.

The target measures retrieval sufficiency, not generator faithfulness. A complete evidence path can still be misread by a generator, and an incomplete path can occasionally support a narrow answer. A production system should pair MAGAF with citation verification, claim-level entailment, and answer-side abstention. Null-query handling also deserves a dedicated model because the fixed retrieval-score rule rejected no null queries in Table 3. Finally, the selected threshold favors recall and produces false positives; applications with higher cost for unsupported answers should choose a stricter policy or add retrieval expansion before acceptance.

6. Conclusion

This paper introduced MAGAF, a metadata-aware graph and attention feature framework for predicting whether a multi-hop RAG context contains a complete evidence path. The method combines retrieval-score summaries with source, category, entity, and date agreement; weighted graph cohesion; estimated hop difficulty; and attention-style measures of score dispersion. Hybrid-MetaGraph retrieval achieved Recall@4 0.501 and CompleteRecall@4 0.320, while the calibrated MAGAF predictor achieved AUROC 0.884, F1 0.765, Brier score 0.119, and ECE 0.072.

The empirical results support two conclusions. First, ordinary relevance and complete-path retrieval are materially different objectives: many queries recover some evidence while missing a required bridge document. Second, metadata, graph structure, and attention-style score features provide complementary information for estimating that gap. The resulting probability can serve as a pre-generation guardrail that accepts strong contexts, expands uncertain retrievals, and abstains when evidence is unlikely to be complete. This decision layer offers a practical way to reduce unsupported multi-hop generation without tying retrieval-quality control to a specific language model.

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 Proc. NeurIPS, 2020.

[3] V. Karpukhin et al., “Dense passage retrieval for open-domain question answering,” in Proc. EMNLP, 2020, pp. 6769-6781.

[4] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. EMNLP-IJCNLP, 2019, pp. 3982-3992.

[5] N. Thakur et al., “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” in Proc. NeurIPS Datasets and Benchmarks, 2021.

[6] Z. Yang et al., “HotpotQA: A dataset for diverse, explainable multi-hop question answering,” in Proc. EMNLP, 2018, pp. 2369-2380.

[7] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, “Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions,” in Proc. ACL, 2023.

[8] 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

[9] H. Liu et al., “HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation,” arXiv:2502.12442, 2025.

[10] J. Lin and J. Liu, “Optimizing multi-hop document retrieval through intermediate representations,” arXiv:2503.04796, 2025.

[11] M. Besta et al., “Multi-Head RAG: Solving multi-aspect problems with LLMs,” arXiv:2406.05085, 2024.

[12] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017, pp. 5998-6008.

[13] P. Velickovic et al., “Graph attention networks,” in Proc. ICLR, 2018.

[14] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. ICLR, 2017.

[15] M. Schlichtkrull et al., “Modeling relational data with graph convolutional networks,” in Proc. ESWC, 2018, pp. 593-607.

[16] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” Comput. Netw. ISDN Syst., vol. 30, nos. 1-7, pp. 107-117, 1998.

[17] O. Khattab and M. Zaharia, “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT,” in Proc. SIGIR, 2020, pp. 39-48.

[18] S. Hofstatter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury, “Efficiently teaching an effective dense retriever with balanced topic aware sampling,” in Proc. SIGIR, 2021, pp. 113-122.

[19] J. Johnson, M. Douze, and H. Jegou, “Billion-scale similarity search with GPUs,” IEEE Trans. Big Data, vol. 7, no. 3, pp. 535-547, 2021.

[20] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[21] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001.

[22] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in Advances in Large Margin Classifiers. Cambridge, MA, USA: MIT Press, 1999, pp. 61-74.

[23] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proc. ICML, 2005, pp. 625-632.

[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. ICML, 2017, pp. 1321-1330.

[25] K. Jarvelin and J. Kekalainen, “Cumulated gain-based evaluation of IR techniques,” ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422-446, 2002.

[26] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[27] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333-389, 2009.

[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,” JACS, vol. 3, no. 1, pp. 49-65, Jan. 2023, doi: 10.69987/JACS.2023.30104.

[29] K. Zhang, S. Meng, and E. Zhou, “Evidence-grounded trading desk risk memos over SEC filings: Retrieval-augmented generation with XBRL numeric verification,” JACS, vol. 3, no. 2, pp. 60-76, Feb. 2023, doi: 10.69987/JACS.2023.30205.

[30] Q. Wu, J. Bai, and X. Zhou, “Evidence-grounded financial RAG: Reducing numerical hallucination in LLM-generated corporate risk memos,” JACS, vol. 3, no. 3, pp. 65-84, Mar. 2023, doi: 10.69987/JACS.2023.30306.

[31] J. Nie and D. Zheng, “Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal,” JACS, vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.

[32] Y. Li, "Execution-feedback and retrieval-augmented generation for conversational text-to-SQL: From one-shot questions to clarification-driven executable dialogs," JACS, vol. 3, no. 2, pp. 1-17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[33] B. Zhang, H. Rao, and D. Zhao, "Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents," JACS, vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[34] 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," JACS, vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[35] C. Li, J. Bai, and S. Wang, "Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications," JACS, vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[36] S. Zhou, Z. Li, and E. Wang, "Long-document RAG for contractual and insurance clause analysis in receivables RWA structures," JACS, vol. 4, no. 8, pp. 88-104, Aug. 2024, doi: 10.69987/JACS.2024.40810.

[37] S. Chen, W. Su, and J. Ma, "Self-correcting text-to-SQL agent with error feedback: A reproducible closed-loop evaluation on compact executable SQLite benchmarks," JACS, vol. 4, no. 11, pp. 86-104, Nov. 2024, doi: 10.69987/JACS.2024.41107.

[38] Y. Li, "Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset," JACS, vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[39] W. Su, S. Chen, and C. Zhao, "Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[40] Y. Chen and H. Xu, "Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141-164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[41] W. Su, S. Chen, and E. Qian, "Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327-340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[42] Q. Xin, "Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation," AVITEC, vol. 8, no. 2, p. 247, May 2026, doi: 10.28989/avitec.v8i2.3974.

[43] Z. Li, K. Zhang, and A. Wong, "Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75-90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.