Question Decomposition and Evidence Chain Scoring for Complex RAG Retrieval Quality Evaluation on the FRAMES Benchmark

Ryan Bennett
Electrical Engineering and Computer Science, Purdue University, West Lafayette, IN, USA
ryanb_purdue2024@gmail.com

DOI: 10.63575/CIA.2026.40118

Abstract

Complex retrieval-augmented generation (RAG) systems can retrieve one highly relevant page while still missing the other pages required to support a multi-hop answer. This paper presents QD-ECS, an article-level evaluation procedure that combines deterministic question decomposition with evidence-chain scoring on the FRAMES benchmark. The evaluation covers all 824 FRAMES questions and a closed candidate pool of 2,480 normalized Wikipedia article identifiers derived from the benchmark links. Five retrieval settings are compared: Random, TF-IDF, BM25, BM25-Decomp, and Hybrid-Chain. Retrieval quality is measured with article Recall@k, ChainExact@k, step evidence coverage (SEC@k), NDCG@k, and MRR. Hybrid-Chain achieved the highest Recall@5 (0.499), SEC@5 (0.438), NDCG@5 (0.552), and MRR (0.824), whereas TF-IDF obtained the highest ChainExact@5 (0.180), exceeding Hybrid-Chain by 0.001. The central finding is that first-hit success substantially overstates evidence sufficiency: Hybrid-Chain recovered every required article in the top five for 147 of 824 questions, and its ChainExact@5 declined from 0.373 on two-article questions to 0.011 on questions requiring five or more articles. QD-ECS therefore separates early partial recovery from complete chain recovery and provides a compact diagnostic framework for evaluating multi-hop RAG retrieval before answer generation

Keywords: *retrieval-augmented generation; RAG evaluation; FRAMES benchmark; multi-hop retrieval; question decomposition; evidence-chain scoring; BM25; TF-IDF*

1. Introduction

Retrieval-augmented generation links language models to external collections so that answers can be conditioned on evidence rather than on model parameters alone. The original RAG formulation established the value of non-parametric memory for knowledge-intensive generation [2], while dense passage retrieval demonstrated that learned retrievers can improve open-domain question answering over traditional sparse baselines [3]. As RAG systems have moved from single-fact lookup toward multi-document reasoning, however, a different evaluation question has become increasingly important: did the retriever recover the complete set of sources needed to justify the answer?

This question is especially consequential for multi-hop queries. A system may place one relevant article at rank one and still omit a second or third article that supplies an intermediate entity, a temporal constraint, or a numerical value. A fluent generator can conceal this omission because transformer-based and few-shot language models are capable of producing coherent answers even when the supplied context is incomplete [22], [23]. Surveys of RAG consequently treat retrieval, generation, and evaluation as coupled components rather than independent stages [12]. For retrieval-stage assessment, plausibility is not enough; the evaluator must determine whether the context contains all benchmark-specified evidence.

FRAMES was designed for this end-to-end setting. It combines factuality, retrieval, and reasoning in 824 human-authored questions that require information from multiple Wikipedia articles [1]. Its questions include numerical, temporal, tabular, post-processing, and multiple-constraint reasoning. Earlier benchmarks such as HotpotQA [6], MuSiQue [7], and HybridQA [21] established the importance of compositional question answering, while

MultiHop-RAG provided a retrieval-focused benchmark for multi-hop queries [24]. FRAMES adds a useful article-level supervision signal because each question is associated with a list of relevant Wikipedia links. Those links make it possible to evaluate whether an entire evidence set is present without relying on a generated-answer judge.

Conventional metrics capture only part of this requirement. MRR reports how early the first relevant item appears. Recall@k gives partial credit for recovering some members of a gold set. NDCG@k adds rank sensitivity. None of these metrics alone states whether every required article has been retrieved. For a question that depends on five articles, retrieving one article at rank one can yield an excellent reciprocal rank while leaving most of the reasoning chain unsupported. A strict chain-completeness measure and a smoother step-weighted measure are therefore needed alongside standard ranking metrics.

This paper develops QD-ECS, a deterministic question-decomposition and evidence-chain scoring framework for article-level retrieval evaluation on FRAMES. The method splits each prompt into clause-level and entity-hint queries, compares five transparent retrieval configurations, and evaluates ranked article identifiers with complementary partial-coverage, full-chain, and rank-sensitive metrics. The study is intentionally focused on retrieval rather than answer generation. This separation makes errors auditable: a low chain score identifies missing sources before the generator can obscure the retrieval failure with a plausible response.

The contributions are threefold. First, QD-ECS defines an article-level protocol that treats every normalized FRAMES gold link as a required evidence step. Second, it reports a complete comparison over all 824 questions and a 2,480-article candidate pool using fixed retrieval settings and paired bootstrap analysis. Third, it characterizes how chain completeness changes with evidence-set length and reasoning type, thereby revealing failure patterns that first-hit metrics do not expose.

2. Literature Review

2.1 Retrieval foundations and multi-hop benchmarks

The retrieval component of QD-ECS draws on a long line of information-retrieval research. Vector-space retrieval represents queries and documents through weighted lexical features [5], and BM25 provides a probabilistic ranking function with term-frequency saturation and document-length normalization [4]. These methods remain valuable evaluation baselines because their scores are transparent and their behavior can be inspected at the token level. Dense Passage Retrieval later showed that dual-encoder representations can recover semantically related passages that do not share exact lexical forms [3]. The present study does not claim that sparse retrieval is universally strongest; it uses sparse methods to make the behavior of the proposed chain metrics easy to trace.

Question-answering benchmarks have progressively increased the complexity of the retrieval target. Natural Questions evaluates answers derived from real search queries [8], TriviaQA uses large-scale distant supervision [9], and ELI5 emphasizes long-form explanatory answers [10]. HotpotQA introduced diverse, explainable multi-hop questions with supporting facts [6], while MuSiQue constructed multi-hop questions through controlled composition to reduce shortcut solutions [7]. HybridQA combined tabular and textual evidence [21]. These benchmarks established that evidence can be distributed across documents and modalities, but their supervision formats and retrieval settings differ from the article-link structure available in FRAMES.

A further distinction concerns whether a benchmark genuinely requires multi-step retrieval. Analyses of compositional questions have shown that some nominally multi-hop examples can be solved through shortcuts or single-hop associations [18]. MultiHop-RAG directly targets retrieval for multi-hop queries and highlights the difficulty of assembling evidence across documents [24]. FRAMES complements that line of work by organizing questions around practical RAG failure modes and by supplying article links that can be normalized into an explicit evidence set [1]. This combination motivates a scoring protocol that measures both partial article recovery and exact chain completion.

2.2 Question decomposition and iterative retrieval

Question decomposition aims to expose intermediate information needs that are hidden inside a complex prompt. Chain-of-thought prompting showed that explicit intermediate reasoning can improve multi-step problem solving [13], and ReAct integrated reasoning traces with external actions such as search [14]. Work on the compositionality gap further demonstrated that models often know the component facts of a question without reliably composing them into a final answer [15]. Self-RAG extended this idea by learning when to retrieve, generate, and critique through self-reflection [16]. Together, these studies suggest that decomposition can improve candidate discovery, but they also warn that local steps must remain connected to the global constraints of the original query.

Related systems in executable and code-centered domains reinforce this point. Execution-feedback RAG for conversational text-to-SQL uses database results and clarification turns to refine a query plan [29], while a self-correcting text-to-SQL agent feeds execution errors back into a closed repair loop [33]. In code intelligence, retrieval-summary integration has been used to make retrieved code both findable and explainable [34]. A budgeted multi-hop retrieval agent similarly treats compositional question answering as a sequence of retrieval decisions under a finite search budget [35]. These approaches differ in application, but they share a structural insight: intermediate queries are useful when their outputs are checked against an overarching objective.

QD-ECS adopts a deliberately lightweight version of this principle. It creates deterministic clause and entity-hint queries rather than relying on a stochastic language-model planner. The original prompt is always retained, so decomposition broadens candidate discovery without replacing global context. The comparison between BM25-Decomp and Hybrid-Chain tests this design choice directly: decomposition is evaluated both as a primary retrieval driver and as an auxiliary signal.

2.3 Evidence sufficiency, factuality, and provenance

Evidence-chain evaluation is related to, but distinct from, answer factuality. TruthfulQA shows that linguistically convincing answers may reproduce common falsehoods [11]. FEVER requires systems to retrieve evidence for factual claims [20], and ERASER separates a model prediction from the rationale that supports it [19]. RAGAS provides automated measures for retrieval-augmented generation, including dimensions associated with context and answer quality [17]. These lines of research motivate an evaluator that treats the retrieved evidence list as an object of analysis rather than inferring retrieval quality from the final prose alone.

Recent applied work has placed greater emphasis on evidence consistency and traceability. A lightweight hallucination firewall combines evidence checks, self-verification, and a smaller detector to screen enterprise LLM outputs [25]. Evidence-Chain Reliable RAG adds word-level hallucination detection, source attribution, and provenance explanation [31]. Both approaches address generation-stage reliability. QD-ECS operates one stage earlier: it asks whether the benchmark-specified source chain entered the context at all. The two perspectives are complementary because provenance mechanisms cannot cite an article that was never retrieved.

This distinction also clarifies the role of strict and graded metrics. ChainExact@k is analogous to a sufficiency condition: every required source must be present. Recall@k and NDCG@k measure partial progress. SEC@k adds an evidence-step discount so that early recovery of several gold articles scores more highly than late or incomplete recovery. Reporting these metrics together prevents a strong first hit from being mistaken for a fully supported chain.

2.4 Domain-grounded RAG and auditable interfaces

Domain applications illustrate why evidence completeness matters beyond benchmark leaderboards. Financial systems have used retrieval and numeric verification to produce trading-desk risk memos grounded in SEC filings [26], and related work has focused specifically on reducing numerical hallucination in corporate risk summaries [27]. Evidence-grounded question answering for tokenized trade-receivable disclosures extends the same concern to capital-market standards [28]. In each case, a plausible narrative is insufficient if a filing, note, or calculation source is absent from the retrieved context.

Operational RAG systems face a similar challenge. Evidence-grounded RAG for cloud-native DevOps emphasizes hallucination-resistant question answering over private operations documents [30], while long-document RAG for contractual and insurance clause analysis must preserve relevant clauses across extended documents [32]. These settings make retrieval failure costly: a missing runbook step, contract exception, or policy clause can change the recommended action. Article-level chain scoring is coarser than clause-level verification, but it provides a reproducible way to study the same completeness principle on a public benchmark.

Human-facing systems add a communication dimension. Trust-calibrated multilingual RAG for humanitarian information platforms combines retrieval evaluation with answerability and uncertainty communication [36]. Numerical-reasoning guardrails for a quantitative research assistant emphasize evidence checks over SEC and FRED data [37]. Evidence-card interfaces for scientific search make source hierarchy and retrieval transparency visible to users [38]. These studies suggest that trustworthy RAG requires three connected layers: source discovery, evidence sufficiency, and transparent presentation. QD-ECS concentrates on the first two layers by producing question-level coverage diagnostics that can later be surfaced in user interfaces or used to trigger an additional retrieval pass.

The literature therefore leaves a clear evaluation gap. Existing benchmarks, RAG evaluators, agentic retrieval methods, and provenance systems address related aspects of complex answering, but ordinary retrieval summaries still tend to emphasize the first relevant result or average relevance. QD-ECS addresses this gap with an article-level protocol that explicitly separates early partial recovery from complete evidence-chain recovery on FRAMES.

3. Method

3.1 QD-ECS overview

QD-ECS evaluates retrieval in five stages: benchmark-question ingestion, deterministic decomposition, closed-pool article representation, method-specific ranking, and evidence-chain scoring. Figure 1 summarizes the workflow. Each normalized gold Wikipedia article is treated as one required evidence step. A retrieved set is fully sufficient at rank k only when every required article appears within the top-k list.

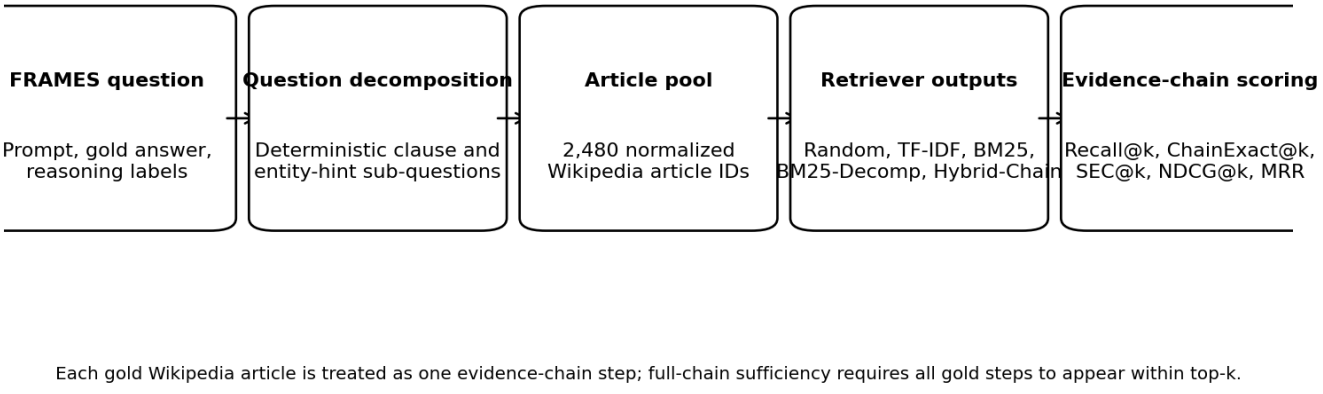

Figure 1. QD-ECS pipeline for deterministic decomposition, article retrieval, and evidence-chain scoring.

The framework is article-ID based. It evaluates whether the retriever recovers the correct Wikipedia pages, not whether it extracts the exact supporting sentence or produces the final answer. This granularity matches the public FRAMES fields used in the experiment, which include prompts, answers, reasoning labels, and Wikipedia links but do not include sentence-level rationales or full article bodies.

3.2 Dataset processing and benchmark profile

The experiment used the public FRAMES test.tsv file and evaluated all 824 records. The wiki_links field served as the primary source of gold evidence links, with individual wikipedia_link columns used as a fallback. URLs were converted to article-level identifiers by standardizing mobile links, converting redirect-style title parameters into /wiki/ paths, removing fragments and query strings, decoding titles, and deduplicating links within each question while preserving their original order.

After normalization, the benchmark contained 2,480 unique article identifiers. Table 1 reports the principal dataset and candidate-pool characteristics. The average question requires 3.198 gold articles, with a median of three and a range from two to eleven.

Table 1. Dataset and article-pool characteristics measured from FRAMES test.tsv.

Dataset or corpus item	Measured value
Questions	824
Unique normalized Wikipedia article IDs	2,480
Mean gold articles per question	3.198
Median gold articles per question	3.0
Minimum / maximum gold articles	2 / 11
Mean prompt words	27.609
Mean answer words	5.180
Input file	data/frames_test.tsv

Table 2 gives the full distribution of evidence-set lengths, and Figure 2 visualizes the same counts. Two- and three-article questions account for most of the benchmark, but 92 questions require five or more articles. This long tail is important because a fixed top-five budget becomes increasingly restrictive as the number of required sources approaches or exceeds k.

Table 2. Evidence-chain length distribution after URL normalization.

Gold articles	Questions	Percent
2	311	37.743
3	287	34.830
4	134	16.262
5	37	4.490
6	19	2.306
7	13	1.578
8	6	0.728
9	3	0.364
10	3	0.364
11	11	1.335

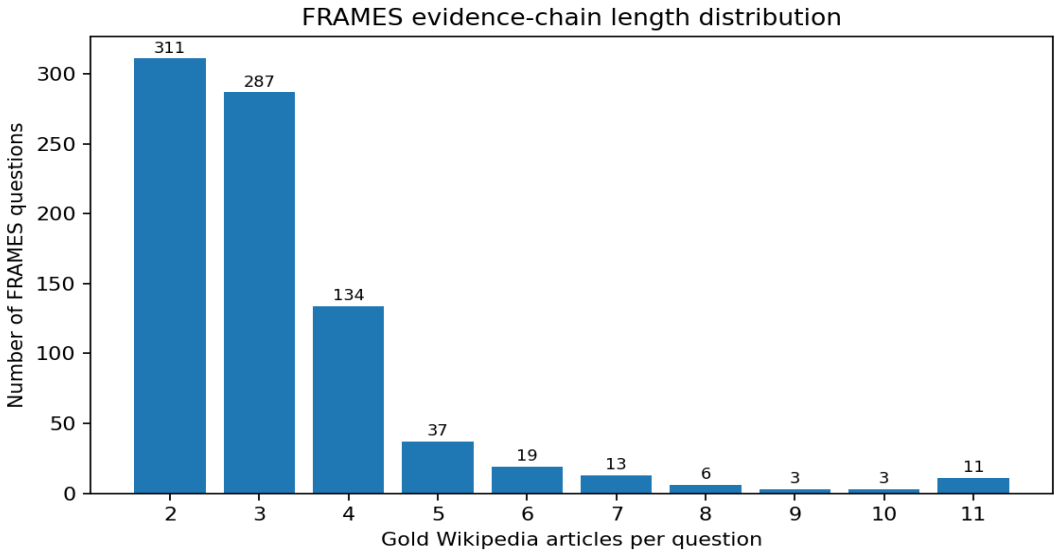

Figure 2. Distribution of normalized gold Wikipedia articles per FRAMES question.

FRAMES reasoning labels are multi-label rather than mutually exclusive. Table 3 and Figure 3 show that multiple constraints are the most common label, followed by numerical, temporal, tabular, and post-processing reasoning. The percentages sum to more than 100 because a question may receive several labels.

Table 3. Multi-label reasoning-type distribution.

Reasoning type	Questions	Percent
Numerical reasoning	293	35.6
Tabular reasoning	236	28.6
Multiple constraints	549	66.6
Temporal reasoning	278	33.7
Post-processing	107	13.0

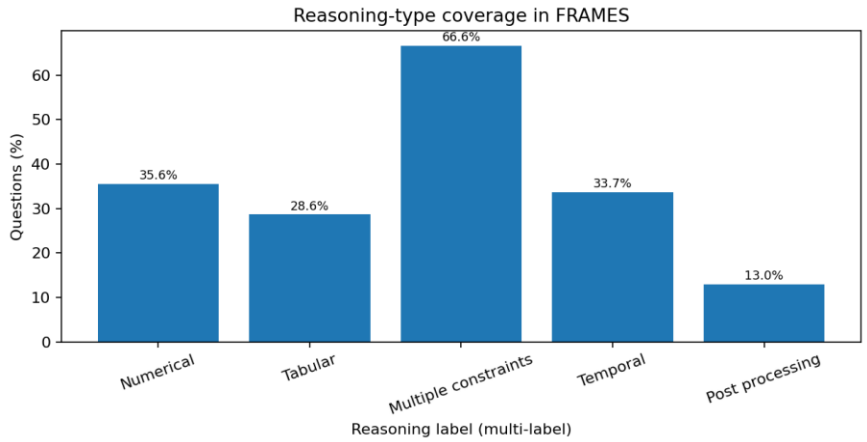

Figure 3. Coverage of FRAMES reasoning labels; labels are multi-label, so percentages sum above 100%.

3.3 Deterministic question decomposition

The decomposition function is deterministic so that repeated runs produce identical sub-questions. Each prompt is split at sentence boundaries, semicolons, colon-like delimiters, and selected logical or wh-phrase boundaries, including which, who, where, when, what, how, after, before, and whose. Clauses containing at least three non-stopword tokens are retained. Capitalized entity spans are also extracted and combined into a compact entity-hint query. The complete original prompt is appended as the final query in every decomposition.

Retaining the original prompt is essential. A clause may contain a relation such as first, same, or after without the entity that makes the relation meaningful. Treating that clause as an independent search query can promote generic distractors. QD-ECS therefore uses decomposition in two ways: BM25-Decomp makes the decomposed queries the principal retrieval inputs, whereas Hybrid-Chain treats them as auxiliary signals added to full-prompt retrieval. This contrast isolates the benefit and risk of decomposition under the same candidate pool.

3.4 Article-pool construction and representation

Every unique normalized gold article appearing anywhere in FRAMES becomes a candidate document for every question. The resulting 2,480-document pool is closed but cross-question: a question is ranked against its own gold articles and against thousands of distractors that are relevant to other FRAMES questions. This construction avoids a question-specific candidate set that would make the target articles artificially easy to identify.

Because the public file supplies article identifiers rather than article bodies, each candidate is represented through the normalized article title and URL-path tokens. The title is repeated three times to emphasize the primary identifier. For example, the URL ending in /wiki/Punxsutawney_Phil becomes the token sequence Punxsutawney Phil. Fragments are removed so that multiple links to sections of the same page map to one article identifier. The resulting task is title-level article retrieval within the benchmark pool, not passage retrieval over a full Wikipedia snapshot.

3.5 Retrieval settings

Five retrieval settings were evaluated with the fixed parameters listed in Table 4. Random provides a deterministic sanity check. TF-IDF follows the vector-space tradition [5] with word unigrams and bigrams, sublinear term frequency, and L2 normalization. BM25 uses k1 = 1.2 and b = 0.75 following the probabilistic relevance framework [4]. BM25-Decomp applies BM25 to each decomposed query and combines the ranked lists through reciprocal-rank fusion [39]. Hybrid-Chain combines full-prompt TF-IDF and BM25 with normalized decomposition and entity signals.

Table 4. Retrieval settings and fixed parameters.

Method	Retriever or evaluator setting	Fixed parameterization
Random	Deterministic shuffle of the 2,480-article pool	seed = 13; MD5 prompt hash
TF-IDF	Full prompt matched to title and URL-token documents	word 1-2 grams; sublinear TF; L2 norm
BM25	Full prompt matched to title and URL-token documents	k1 = 1.2; b = 0.75
BM25-Decomp	Clause and entity queries ranked separately and fused	reciprocal-rank fusion; C = 60
Hybrid-Chain	Full-prompt lexical scores plus decomposition and entity signals	*0.62 TF-IDF + 0.28 BM25 + 0.24 decomposition + 0.06 entity*

The Hybrid-Chain weights were fixed for the complete evaluation and were not tuned per question. Full-prompt signals remain dominant, while decomposition is strong enough to elevate candidate evidence steps that appear in only one clause. The entity term provides a small additional preference for title matches to named entities extracted from the prompt.

3.6 Evidence-chain metrics

Let Gq denote the set of m normalized gold article identifiers for question q, and let Rq,k denote the top-k retrieved identifiers. Article Recall@k measures the fraction of required articles recovered within the retrieval budget:

Recall@k(q) = |Gq ∩ Rq,k| / |Gq|.

ChainExact@k is a strict sufficiency indicator. It equals one only when the entire gold set is contained in the top-k ranking:

$$\text{ChainExact@k}(q) = 1[G_q \subseteq R_{q,k}].$$

SEC@k is a graded step evidence coverage score. For each gold article found within the top-k list, the score applies an exponential rank discount; a missing article contributes zero:

$$\text{SEC@k}(q) = (1 / |G_q|) \sum_{g \in G_q} 1[\text{rank}_q(g) \leq k] \exp(-(\text{rank}_q(g) - 1) / k).$$

NDCG@k is computed with binary article relevance and the standard logarithmic rank discount [40]. MRR is the reciprocal rank of the first retrieved gold article. The reported values are macro-averages over all 824 questions. Together, the metrics answer different questions: Recall measures partial coverage, ChainExact tests full sufficiency, SEC evaluates early step recovery, NDCG captures rank-sensitive relevance, and MRR records the first successful hit.

3.7 Experimental protocol

Every method ranked the complete 2,480-document pool for every question. The main analysis reports recall at k = 2, 3, 5, and 10; ChainExact and SEC at k = 5 and 10; and NDCG@5 and MRR. Question-level rankings and decompositions were retained so that aggregate scores could be traced to individual examples. The random seed was fixed at 13, and the random baseline used a stable MD5-based prompt hash rather than a process-dependent language hash.

Paired uncertainty estimates were obtained with 1,000 bootstrap resamples over question identifiers using seed 13. Each resample preserved the pairing between methods, so the reported interval reflects the distribution of per-question score differences. This design is especially useful for the comparison between TF-IDF and Hybrid-Chain, whose aggregate scores are close.

4. Results and Discussion

4.1 Benchmark profile and evidence burden

The benchmark profile in Tables 1-3 establishes the scale of the retrieval problem. The mean evidence set contains 3.198 articles, but the maximum is eleven. Two- and three-article questions dominate, yet 92 questions require five or more sources. Figure 2 shows this long tail clearly. The reasoning distribution in Figure 3 also indicates that retrieval must often satisfy several constraints at once: 66.6% of questions carry the multiple-constraints label, and substantial subsets require numerical, temporal, or tabular reasoning.

These characteristics make a first-hit metric insufficient. An early relevant article may identify the topic while leaving an entity bridge, date constraint, list page, or numerical source unretrieved. The evidence burden grows with the number of required articles, and the top-k budget becomes a structural limit when the gold set approaches k.

4.2 Main retrieval comparison

Table 5 reports the complete comparison. Hybrid-Chain achieved the highest measured Recall@5 (0.499), SEC@5 (0.438), NDCG@5 (0.552), and MRR (0.824). TF-IDF was a close second on those metrics and obtained the highest ChainExact@5 (0.180), compared with 0.178 for Hybrid-Chain. BM25 remained competitive but was consistently below TF-IDF on the title-only article pool. BM25-Decomp was substantially weaker, indicating that decomposition cannot replace the complete prompt when article titles are the only document content.

Table 5. Main retrieval and evidence-chain results on 824 FRAMES questions.

Method	R@2	R@3	R@5	R@10	CE@5	CE@10	SEC@5	NDCG@5	MRR
Random	0.001	0.001	0.002	0.003	0.000	0.000	0.001	0.002	0.004
TF-IDF	0.401	0.451	0.496	0.543	0.180	0.223	0.434	0.547	0.816

Method	R@2	R@3	R@5	R@10	CE@5	CE@10	SEC@5	NDCG@5	MRR
BM25	0.395	0.442	0.489	0.538	0.176	0.218	0.427	0.535	0.795
BM25-Decomp	0.220	0.247	0.275	0.322	0.076	0.101	0.240	0.297	0.468
Hybrid-Chain	0.406	0.452	0.499	0.544	0.178	0.222	0.438	0.552	0.824

R@k = article Recall@k; CE@k = ChainExact@k; SEC@k = step evidence coverage score.

Figure 4 visualizes the four top-five metrics that most directly compare relevance and sufficiency. The Random baseline is near zero across all measures, providing a sanity check for the 2,480-document pool. The strongest lexical systems recover roughly half of the required articles at top five, but their ChainExact@5 remains below 0.20. This gap is the central empirical distinction between partial relevance and complete evidence support.

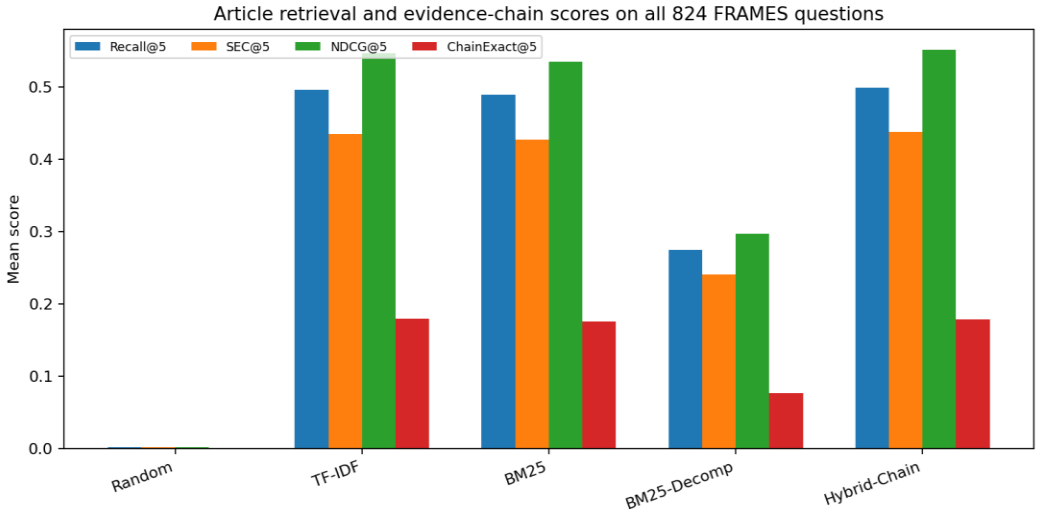

Figure 4. Comparison of Recall@5, SEC@5, NDCG@5, and ChainExact@5 across the five retrieval settings.

Figure 5 shows Recall@k at k = 2, 3, 5, and 10. TF-IDF, BM25, and Hybrid-Chain remain close across the full range, with Hybrid-Chain holding a small advantage at most cutoffs. BM25-Decomp improves as k grows but does not approach the full-prompt systems. The result supports the use of decomposition as a secondary candidate-discovery signal rather than as a substitute for the original query.

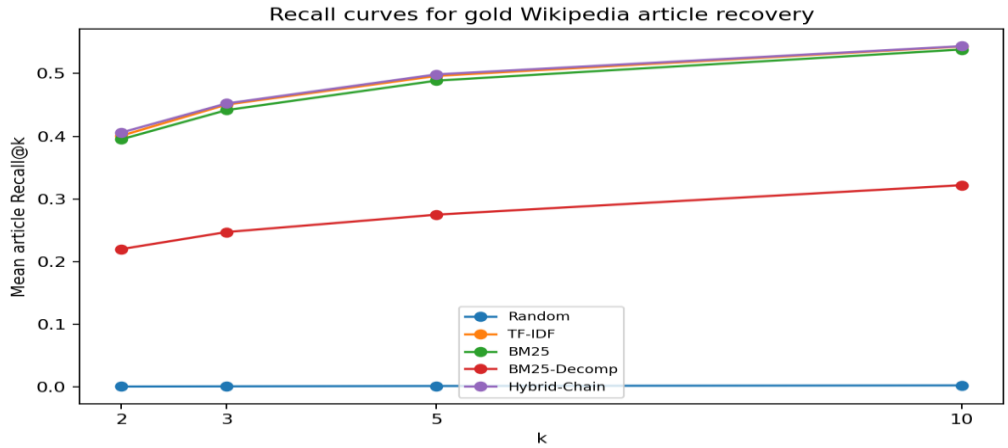

Figure 5. Recall@k curves for normalized gold Wikipedia article recovery.

The difference between TF-IDF and BM25 is consistent with the document representation. Candidate documents are short title strings with repeated title tokens. Cosine-normalized unigram and bigram overlap is therefore highly

informative, whereas BM25 document-length normalization has limited variation to exploit. This finding is specific to the closed title-level pool and should not be generalized to full Wikipedia passages.

4.3 Evidence-chain length and exact recovery

Table 6 stratifies the results by the number of gold articles. For Hybrid-Chain, Recall@5 declines from 0.633 on two-article questions to 0.290 on questions requiring five or more articles. The decline is much sharper for ChainExact@5: 0.373 for two articles, 0.087 for three, 0.037 for four, and 0.011 for five or more. Figure 6 shows that exact recovery becomes rare as the required chain grows.

Table 6. Performance by evidence-chain length bucket.

Method	Gold-article bucket	R@5	CE@5	SEC@5
TF-IDF	2	0.635	0.379	0.565
TF-IDF	3	0.470	0.080	0.409
TF-IDF	4	0.369	0.037	0.322
TF-IDF	5+	0.292	0.022	0.234
BM25	2	0.627	0.373	0.559
BM25	3	0.467	0.084	0.403
BM25	4	0.366	0.037	0.316
BM25	5+	0.269	0.000	0.215
BM25-Decomp	2	0.392	0.174	0.349
BM25-Decomp	3	0.254	0.031	0.220
BM25-Decomp	4	0.159	0.000	0.135
BM25-Decomp	5+	0.113	0.000	0.089
Hybrid-Chain	2	0.633	0.373	0.568
Hybrid-Chain	3	0.481	0.087	0.417
Hybrid-Chain	4	0.368	0.037	0.321
Hybrid-Chain	5+	0.290	0.011	0.235

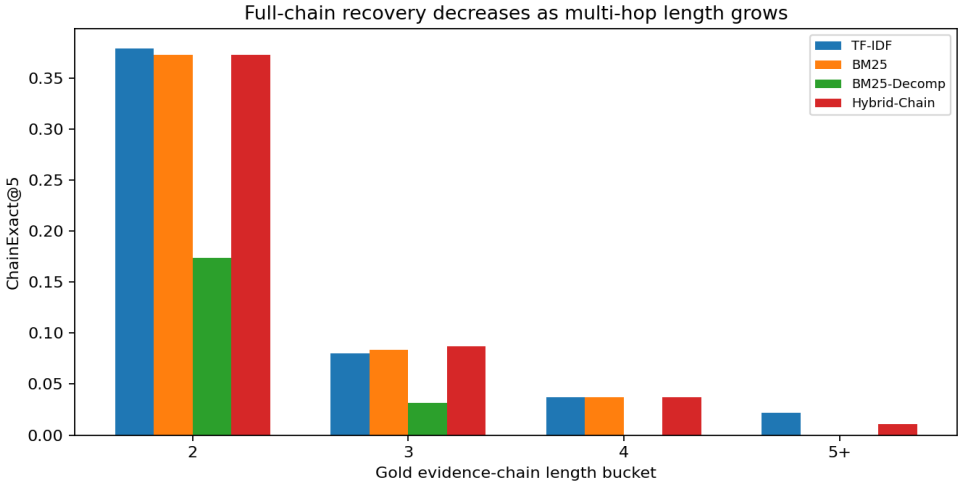

Figure 6. ChainExact@5 by gold evidence-chain length bucket.

The pattern is partly mathematical and partly semantic. A top-five list has less room for distractors when the gold set contains four or more articles. At the same time, longer chains tend to include intermediate entities that are not stated directly in the original question. A larger context budget helps only modestly: Hybrid-Chain rises from

ChainExact@5 = 0.178 to ChainExact@10 = 0.222. Doubling k recovers additional complete chains, but most questions remain incomplete.

This result aligns with the motivation behind multi-step and budgeted retrieval systems [24], [35]. A production retriever cannot rely solely on a wider first-pass ranking. It needs to identify which evidence steps are already covered, infer what remains missing, and issue targeted follow-up queries. ChainExact and SEC provide complementary signals for that decision: the former identifies complete success, while the latter indicates how much useful evidence has already been placed early in the ranking.

4.4 Performance by reasoning type

Table 7 reports Hybrid-Chain performance across the multi-label reasoning categories. Numerical reasoning has the strongest Recall@5 (0.553), ChainExact@5 (0.259), SEC@5 (0.481), and NDCG@5 (0.601). Multiple-constraint questions are more difficult, with Recall@5 = 0.435 and ChainExact@5 = 0.095. Temporal, tabular, and post-processing questions fall between these extremes.

Table 7. Hybrid-Chain performance by reasoning type.

Reasoning type	R@5	CE@5	SEC@5	NDCG@5
Multiple constraints	0.435	0.095	0.381	0.498
Numerical reasoning	0.553	0.259	0.481	0.601
Post-processing	0.469	0.150	0.407	0.532
Tabular reasoning	0.468	0.165	0.408	0.516
Temporal reasoning	0.493	0.158	0.436	0.550

Figure 7 compares SEC@5 by method and reasoning type. The full-prompt systems are strongest across all categories, while BM25-Decomp is consistently lower. Multiple-constraint questions remain the most difficult because relevant titles may depend on an intersection of entities and relations rather than on one explicit title cue. The result shows that FRAMES reasoning labels are useful retrieval diagnostics, not only answer-generation labels.

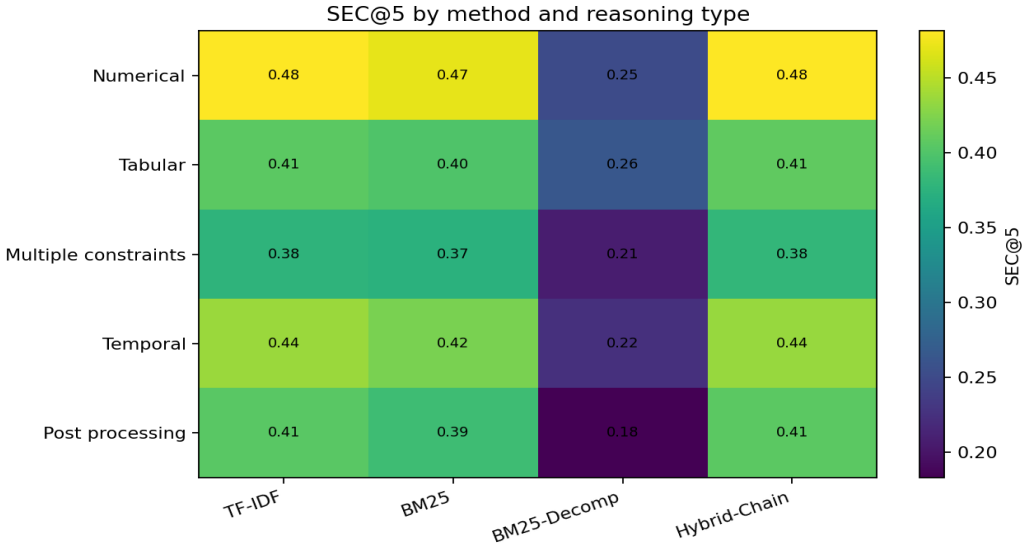

Figure 7. SEC@5 by retrieval method and FRAMES reasoning type.

The comparatively strong numerical category should not be interpreted as evidence that numerical reasoning itself is easy. The retrieval task scores whether the relevant article pages are present; it does not verify arithmetic or table extraction. Domain systems that perform numerical checks over financial records [26], [27], [37] address a later stage of the pipeline. QD-ECS identifies whether the required source pages are available to such a verifier.

4.5 Paired uncertainty analysis

Table 8 reports paired bootstrap differences for Hybrid-Chain against the three non-random baselines. Relative to BM25, Hybrid-Chain improves Recall@5 by 0.010 with a 95% interval of [0.003, 0.017], SEC@5 by 0.011 with [0.006, 0.016], and NDCG@5 by 0.017 with [0.010, 0.024]. The ChainExact@5 difference is small and its interval includes zero.

Table 8. Paired bootstrap differences for Hybrid-Chain against the baselines.

Paired comparison	Metric	Mean difference	Bootstrap 95% CI
Hybrid-Chain minus TF-IDF	Recall@5	0.002	[-0.002, 0.008]
Hybrid-Chain minus TF-IDF	ChainExact@5	-0.001	[-0.008, 0.005]
Hybrid-Chain minus TF-IDF	SEC@5	0.003	[0.000, 0.007]
Hybrid-Chain minus TF-IDF	NDCG@5	0.005	[0.001, 0.010]
Hybrid-Chain minus BM25	Recall@5	0.010	[0.003, 0.017]
Hybrid-Chain minus BM25	ChainExact@5	0.002	[-0.008, 0.012]
Hybrid-Chain minus BM25	SEC@5	0.011	[0.006, 0.016]
Hybrid-Chain minus BM25	NDCG@5	0.017	[0.010, 0.024]
Hybrid-Chain minus BM25-Decomp	Recall@5	0.224	[0.204, 0.243]
Hybrid-Chain minus BM25-Decomp	ChainExact@5	0.102	[0.081, 0.124]
Hybrid-Chain minus BM25-Decomp	SEC@5	0.198	[0.181, 0.214]
Hybrid-Chain minus BM25-Decomp	NDCG@5	0.254	[0.234, 0.274]

Positive values favor Hybrid-Chain. Intervals use 1,000 paired bootstrap resamples over question identifiers.

Against TF-IDF, Hybrid-Chain gains 0.003 SEC@5 and 0.005 NDCG@5, with the NDCG interval entirely above zero. The ChainExact difference is -0.001 with an interval spanning zero. This pattern is informative rather than contradictory: Hybrid-Chain improves the ordering and weighted coverage of evidence steps, but a binary complete-chain score changes only when the final missing article enters the top five. ChainExact is therefore stricter and less sensitive to small ranking improvements.

4.6 Qualitative chain diagnostics

Table 9 presents five representative Hybrid-Chain rankings. The covered-at-five column expresses retrieval support directly as recovered gold articles divided by required gold articles. QID 2 is a complete 2/2 chain: both Punxsutawney Phil and United States Capitol appear in the top five. In contrast, QID 8 retrieves Monaco and Moon but misses the Chevrolet and Apollo-related pages needed to complete the six-article chain.

Table 9. Qualitative evidence-chain examples from Hybrid-Chain.

QID	Covered@5	First decomposed steps	Gold titles (truncated)	Hybrid top titles (truncated)
0	1/5	If my future wife has the same first name as	President of the United States; James	How I Met Your Mother; President of

QID	Covered@5	First decomposed steps	Gold titles (truncated)	Hybrid top titles (truncated)
		the 15th first lady of the United States' mother | her surname is the same as the second assassinated president's mother's maiden name	Buchanan; Harriet Lane; List of presidents of the United States who died in office	the United States; United States; List of How I Met Your Mother episodes
2	2/2	How many years earlier would Punxsutawney Phil have to be canonically alive to have made a Groundhog Day prediction in the same state as the US capitol? | Punxsutawney Phil Groundhog Day	Punxsutawney Phil; United States Capitol	Punxsutawney Phil; Phil Quartararo; A Folk Song A Day; United States Capitol
8	2/6	the largest ward in the country of Monaco | A General Motors vehicle is named after the largest ward in Monaco. How many people had walked on the moon as of the first model year?	Monaco; List of Chevrolet vehicles; Chevrolet Monte Carlo; Moon	Monaco; The Ward (film); Moon; List of largest extant lizards
16	2/3	On March 7, 2012, director James Cameron explored a very deep undersea trench | how many times would the tallest building in San Francisco fit from the bottom of the New Britain Trench?	James Cameron; Solomon Sea; List of tallest buildings in San Francisco	List of tallest buildings in San Francisco; James Cameron; List of tallest buildings in New York City; Billboard Year-End Hot 100 singles of 2016
22	1/5	According to the 1990 United States census | what was the total population of Oklahoma cities with at least 100,000 residents in the 2020 census?	List of United States cities by population; Oklahoma City; Tulsa, Oklahoma; Broken Arrow, Oklahoma	List of United States cities by population; President of the United States; United States; List of presidents of the United States

The examples clarify why decomposition is useful but fragile. Named entities such as Punxsutawney Phil or James Cameron produce strong title matches. Relational fragments such as same first name, largest ward, or cities with at least 100,000 residents require an intermediate mapping that is not visible in a title-only representation.

When the intermediate entity is implicit, the retriever often identifies the topic but not every page needed for the calculation.

Table 10 summarizes the number of missing gold articles in the Hybrid-Chain top-five rankings. Complete coverage occurs for 147 questions. Another 282 questions miss exactly one article, and 212 miss two. These near-complete cases are particularly actionable because a targeted second retrieval pass may close the remaining gap.

Table 10. Hybrid-Chain missing-evidence distribution at top five.

Missing gold articles at top five	Questions
0	147
1	282
2	212
3	101
4	40
5	7
6	17
7	3
8	4
9	4
10	5
11	2

Figure 8 visualizes the same distribution. A high MRR combined with one or two missing articles indicates that the system has found a useful starting point but not a sufficient context. This diagnostic is more operationally informative than a single answer score because it identifies whether the next improvement should target candidate generation, iterative search, or ranking of already discovered pages.

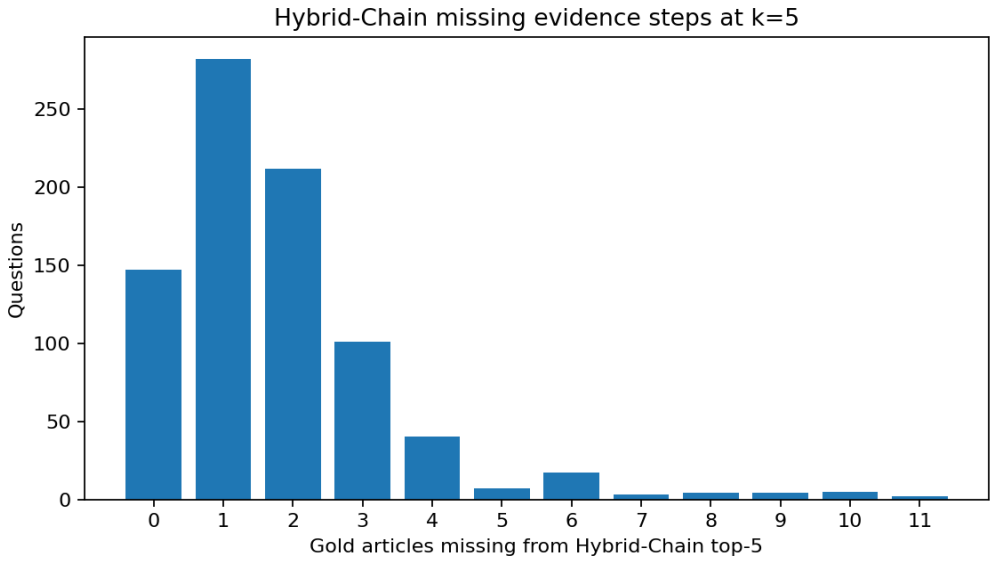

Figure 8. Distribution of gold evidence steps missing from the Hybrid-Chain top-five rankings.

4.7 Implications for complex RAG evaluation

The experiments support three evaluation principles. First, first-hit relevance and evidence sufficiency must be reported separately. Hybrid-Chain reaches MRR = 0.824 while completing only 17.8% of chains at top five. Second, decomposition should augment rather than replace the original query. BM25-Decomp loses global

constraints, whereas Hybrid-Chain preserves them. Third, evidence-set length should be treated as a primary difficulty variable because exact recovery collapses as the number of required sources grows.

These principles connect retrieval evaluation to provenance-oriented RAG. Evidence-chain reliability methods [31], domain-grounded financial and operational systems [26]-[30], and long-document clause analysis [32] all depend on the availability of the correct source set. QD-ECS supplies a retrieval-stage test for that prerequisite. A downstream verifier can then check numerical consistency, sentence-level entailment, or citation quality once the source pages have been retrieved.

The metrics also suggest a practical control loop. A system can begin with a flat full-prompt retriever, add decomposed queries to broaden candidate discovery, and compute SEC@k and ChainExact@k before generation. Low MRR indicates failure to identify even a starting source. High MRR with low SEC indicates that only one part of the chain was found. High SEC with zero ChainExact indicates a near-complete chain that may benefit from one targeted query. Such signals can guide budget allocation in multi-hop agents [35] and can be communicated through transparent evidence displays [38].

Finally, the reasoning-type analysis demonstrates that retrieval evaluation should be stratified. Numerical questions in this title-level setting often contain explicit entity cues, whereas multiple-constraint questions depend more heavily on implicit intersections. Trust-calibrated systems in multilingual and high-stakes environments [36], [37] can use similar stratification to decide when a retrieved context is adequate and when an answer should be deferred pending further evidence.

5. Limitations

The evaluation is conducted at article-ID level. It measures whether a ranked list contains the Wikipedia pages specified by FRAMES, not whether it contains the exact sentence, table cell, infobox field, or paragraph that proves the answer. Passage-level evidence scoring would require article snapshots and supporting-span annotations that are not part of the test.tsv fields used here.

The candidate pool contains 2,480 normalized article identifiers drawn from all FRAMES gold links. It therefore includes realistic cross-question distractors but is much smaller than the full Wikipedia corpus. The measured values characterize closed-pool title retrieval and should not be interpreted as web-scale retrieval performance.

Candidate documents are represented by titles and URL-path tokens because full article bodies are unavailable in the benchmark file. This representation favors explicit title cues and cannot recover an article solely because its body contains an intermediate fact. Dense retrieval over a fixed Wikipedia snapshot may change both absolute performance and the relative ordering of methods.

The decomposition procedure is rule-based. Its deterministic behavior supports consistent comparison, but the resulting clauses are heuristic rather than annotated reasoning steps. A trained semantic parser or a language-model planner could generate more precise sub-questions, although it would also introduce model dependence and additional sources of variability.

The study isolates retrieval and does not score generated answers. Complete article recovery is a necessary evidence condition under the benchmark labels, but it does not guarantee that a generator will extract the correct facts, perform the required calculation, or produce a faithful citation. Conversely, a model with memorized knowledge might answer correctly despite incomplete retrieval. Joint retrieval-generation evaluation remains an important complementary analysis.

Finally, the Hybrid-Chain weights are fixed rather than selected through a separate development split. This choice keeps the comparison uniform across all questions, but it does not establish that the weights are optimal. Future work can evaluate learned fusion, adaptive retrieval budgets, passage expansion, and iterative stopping rules while retaining the same chain-aware metrics.

6. Conclusion

QD-ECS provides an article-level framework for evaluating whether a complex RAG retriever recovers the complete evidence chain required by FRAMES. Across all 824 questions and a 2,480-article candidate pool, Hybrid-Chain achieved the strongest Recall@5, SEC@5, NDCG@5, and MRR, while TF-IDF was effectively tied on exact chain recovery. The difference between MRR = 0.824 and ChainExact@5 = 0.178 shows why first-hit relevance cannot serve as a proxy for evidence sufficiency.

The framework turns that gap into an actionable diagnosis. Recall reports partial coverage, SEC measures early evidence-step recovery, and ChainExact identifies whether the retrieved context contains every required source. Performance declines sharply as evidence sets grow, and decomposition helps only when the full prompt remains the dominant signal. These findings support chain-aware evaluation as a standard component of multi-hop RAG development, particularly for systems that require auditable provenance, numerical verification, or calibrated decisions about when additional retrieval is necessary.

References

[1] S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui, "Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation," in Proc. NAACL, 2025.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "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] A. Singhal, "Modern Information Retrieval: A Brief Overview," IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 35-43, 2001.

[6] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "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, "MuSiQue: Multihop Questions via Single-hop Question Composition," Transactions of the Association for Computational Linguistics, vol. 10, pp. 539-554, 2022.

[8] T. Kwiatkowski et al., "Natural Questions: A Benchmark for Question Answering Research," Transactions of the Association for Computational Linguistics, vol. 7, pp. 453-466, 2019.

[9] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," in Proc. ACL, 2017, pp. 1601-1611.

[10] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "ELI5: Long Form Question Answering," in Proc. ACL, 2019, pp. 3558-3567.

[11] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in Proc. ACL, 2022, pp. 3214-3252.

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

[13] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in Proc. NeurIPS, 2022.

[14] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," in Proc. ICLR, 2023.

[15] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis, "Measuring and Narrowing the Compositionality Gap in Language Models," in Proc. EMNLP, 2023, pp. 5687-5711.

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

[17] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv:2309.15217, 2023.

[18] S. Min, E. Wallace, S. Singh, M. Gardner, H. Hajishirzi, and L. Zettlemoyer, "Compositional Questions Do Not Necessitate Multi-hop Reasoning," in Proc. ACL, 2019, pp. 4249-4257.

[19] J. DeYoung et al., "ERASER: A Benchmark to Evaluate Rationalized NLP Models," in Proc. ACL, 2020, pp. 4443-4458.

[20] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-scale Dataset for Fact Extraction and Verification," in Proc. NAACL-HLT, 2018, pp. 809-819.

[21] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang, "HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data," in Findings of EMNLP, 2020, pp. 1026-1036.

[22] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, 2020, pp. 1877-1901.

[23] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 5998-6008.

[24] 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

[25] 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.

[26] 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.

[27] 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.

[28] S. Zhou, Z. Li, and E. Wang, "Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards," JACS, vol. 3, no. 7, pp. 41-57, Jul. 2023, doi: 10.69987/JACS.2023.30704.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[36] 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, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

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

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

[39] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in Proc. SIGIR, 2009, pp. 758-759.

[40] K. Jarvelin and J. Kekalainen, "Cumulated Gain-Based Evaluation of IR Techniques," ACM Transactions on Information Systems, vol. 20, no. 4, pp. 422-446, 2002.