Coverage-Aware Prediction of Multi-Hop RAG Answer Correctness on FRAMES

Callum Hughes
Computer Science, University of Bristol, Bristol, BST, UK
chughes.codes@yahoo.com

DOI: 10.63575/CIA.2026.40202

Abstract

Multi-hop retrieval-augmented generation (RAG) can retrieve context that appears relevant while omitting an article needed to complete the reasoning chain. This study examines whether that evidence deficit can be detected before answer generation. The evaluation covers all 824 questions in the public FRAMES test file, which provides gold answers, reasoning labels, and relevant Wikipedia links. Because the file identifies articles but does not include article bodies, a closed-world title corpus is constructed from every parsed gold link. Four deterministic retrievers are compared: BM25 over normalized titles, word-level TF-IDF, character-level TF-IDF, and a fixed hybrid. Coverage is defined as the fraction of a question's parsed gold article set returned in the top-k. Strict coverage-gated correctness is one only when the full set is present; it therefore measures retrieval readiness rather than neural-reader accuracy. At k=10, the hybrid reaches mean coverage of 0.547 and strict correctness of 0.225; at k=25, the values rise to 0.590 and 0.273. Performance declines sharply as article count grows. In five-fold cross-validation, logistic regression using observable query and ranking signals predicts strict correctness at k=10 with ROC-AUC 0.796. Adding benchmark hop-count metadata raises ROC-AUC to 0.856. The findings show that incomplete multi-hop evidence is both a central retrieval bottleneck and a predictable risk signal that can guide additional retrieval, decomposition, or abstention before generation.

Keywords: *retrieval-augmented generation; FRAMES; multi-hop question answering; retrieval coverage; strict correctness; BM25; TF-IDF; calibration.*

Introduction

Retrieval-augmented generation combines a parametric language model with evidence retrieved at inference time, reducing the need to encode every fact in model parameters and providing a basis for source-grounded responses [3], [4]. Dense passage retrieval, late interaction, and contrastive retrieval training have improved document discovery for open-domain question answering [5], [6], [7]. Multi-hop questions nevertheless create a distinct failure mode. A ranking can surface one highly plausible page while omitting a second page needed for comparison, temporal ordering, numerical aggregation, or bridge-entity resolution. In such cases, fluent generation cannot compensate reliably for absent evidence.

FRAMES was designed to evaluate factuality, retrieval, and reasoning in a unified RAG setting [1]. Its public test split contains 824 natural-language questions with gold answers, reasoning-type labels, and relevant Wikipedia article links [2]. The explicit article sets make the benchmark particularly useful for studying evidence coverage. Rather than asking only whether a retriever finds at least one relevant page, the benchmark permits a stricter question: has the ranking exposed the complete annotated evidence set needed for the task?

This paper investigates whether a retrieval system's readiness to support a correct multi-hop answer can be predicted before generation. For a question q and retrieval depth k, coverage is the proportion of parsed gold articles present in the top-k ranking. Strict coverage-gated correctness is one when coverage equals one and zero otherwise. This label is intentionally tied to evidence availability. It does not assume that every generator will reason correctly when all articles are present, nor that every annotated article is indispensable for every possible

reader. It provides a reproducible measure of whether retrieval has blocked a complete-evidence answer under the benchmark annotation.

The study contributes a deterministic evaluation pipeline in three parts. First, it parses the full FRAMES test file, including comma-separated links stored in the wikipedia_link_11+ field, and normalizes article identities at page level. Second, it compares BM25, word-level TF-IDF, character-level TF-IDF, and a fixed score-level hybrid over a closed-world corpus of all parsed article titles. Third, it evaluates lightweight classifiers that use query characteristics, reasoning labels, ranking geometry, and title overlap to estimate the probability of strict correctness at k=10. A separate analysis adds annotated hop count to quantify how much predictive value resides in evidence-set size.

The scope is narrower than an end-to-end reader comparison. The released benchmark file contains links but not article body text, so the retrieval study operates on normalized titles. This controlled setting isolates document discovery from passage selection and language-model reasoning. Title retrieval is demanding because many FRAMES prompts describe targets indirectly, yet it remains informative: a system that cannot recover the annotated page identities from the query is unlikely to assemble the full reasoning chain without additional search or decomposition.

A coverage-first analysis also clarifies why first-hit metrics can overstate multi-hop readiness. Mean reciprocal rank rewards an early relevant result, while strict correctness remains zero if any required branch is absent. By connecting these views, the study separates evidence acquisition from downstream reasoning and provides an operational signal for deciding whether a RAG controller should answer, retrieve again, decompose the query, or abstain.

Literature Review

Retrieval for Knowledge-Intensive Question Answering

RAG and REALM established retrieval as a trainable component of knowledge-intensive language modeling [3], [4]. Subsequent work improved retrieval through dense dual encoders, late interaction, and hard-negative contrastive learning [5], [6], [7]. These neural methods are central to modern open-domain systems, but classical lexical models remain useful when documents are represented by names, headings, or short identifiers. TF-IDF emphasizes discriminative term overlap [8], while BM25 adjusts term saturation and document length within a probabilistic relevance framework [9]. Their transparency and deterministic scoring make them suitable baselines for isolating title-discovery behavior.

Generative readers such as Fusion-in-Decoder and retrieval-augmented few-shot models demonstrate that multiple retrieved passages can improve answer synthesis [16], [17]. Their effectiveness, however, depends on the evidence exposed by the retriever. Self-reflective methods add retrieval and critique decisions during generation [18], and broad surveys identify retrieval quality, faithfulness, and evaluation design as persistent RAG challenges [19]. These developments motivate a separate monitor that estimates evidence sufficiency before the reader commits to an answer.

Multi-Hop Benchmarks and Evidence Structure

Multi-hop evaluation has been developed through datasets with complementary structures. HotpotQA emphasizes explainable reasoning across documents [10]; MuSiQue constructs compositional questions from validated single-hop components [11]; QAngaroo targets cross-document reading comprehension [12]; and HybridQA combines tables with linked text [13]. Natural Questions and TriviaQA provide large-scale open-domain supervision [14], [15], while OpenBookQA, GSM8K, and ELI5 emphasize open-book reasoning, mathematical reasoning, and long-form explanation [21], [22], [23]. TruthfulQA focuses on whether models reproduce common falsehoods [20]. FRAMES differs by placing gold answers, reasoning categories, and relevant article sets in one end-to-end RAG benchmark [1], [2], enabling direct analysis of how much annotated evidence a ranking covers.

Most retrieval evaluations report recall at a fixed cutoff or the rank of the first relevant item. Those metrics are appropriate for single-answer document search but can conceal incomplete reasoning chains. In multi-hop tasks,

the unit of success is often a set rather than a single page. Set coverage therefore complements conventional ranking metrics by distinguishing a ranking that retrieves one anchor article from a ranking that supplies every annotated branch.

Evidence Grounding, Provenance, and Hallucination Control

Evidence-aware RAG research increasingly treats grounding as a system property rather than a prompt-level preference. A lightweight hallucination firewall has combined evidence consistency, self-checking, and small-model detection on TruthfulQA [25]. Financial RAG studies have paired retrieval with structured numerical verification over SEC and corporate evidence [26], [27], demonstrating that source presence alone is insufficient when answers depend on exact quantities. Related work on tokenized receivable disclosures and evidence-chain reliability has emphasized source attribution, provenance, and fine-grained hallucination detection [30], [31]. Together, these studies support evaluation designs that make missing or inconsistent evidence visible before a generated statement is accepted.

Operational applications further show that evidence requirements differ by task. Selective refusal has been integrated with retrieval-augmented failure narratives for ambiguous HDFS anomalies [28], while execution feedback has been used to repair conversational text-to-SQL results [29]. Evidence-grounded systems for private DevOps documents and bilingual cloud-incident triage address retrieval under domain vocabulary and access constraints [32], [33]. Retrieval-summary integration for code intelligence and long-document RAG for contractual analysis illustrate the same challenge at different document scales [34], [35]. These systems converge on a common principle: confidence should reflect the quality and completeness of the supporting evidence, not only the fluency of the final response.

Adaptive Retrieval and Coverage-Aware Control

Adaptive retrieval policies attempt to spend additional search effort only when it is useful. Risk-calibrated biomedical search has studied selective query expansion under a coverage-risk trade-off [36], and a budgeted multi-hop retrieval agent has compared fixed, iterative, decomposition, and budget-aware policies on compositional question answering [37]. Evidence ranking has also been incorporated into narrative-aware scientific claim verification [38]. These lines of work suggest that retrieval should be controlled by an explicit estimate of evidence risk. The present study contributes a complementary prediction problem: using only prompt and ranking signals, estimate whether the complete FRAMES article set has already been retrieved.

The resulting perspective connects retrieval evaluation with selective prediction. Rather than treating coverage as a diagnostic reported after generation, the system estimates coverage failure before generation and uses that estimate to choose an action. This is especially relevant for multi-hop questions, where a convincing first result can increase confidence even as the remaining evidence branches stay hidden.

Method

Study Design and Evaluation Pipeline

The experiment evaluates every row in the public FRAMES test file. The benchmark supplies the question, gold answer, one or more reasoning labels, and relevant Wikipedia links. The pipeline proceeds from link parsing and title normalization to closed-world ranking, coverage measurement, strict-label construction, feature extraction, and cross-validated prediction. Figure 1 summarizes these stages and separates observable deployment features from hop-count metadata used only in the secondary analysis.

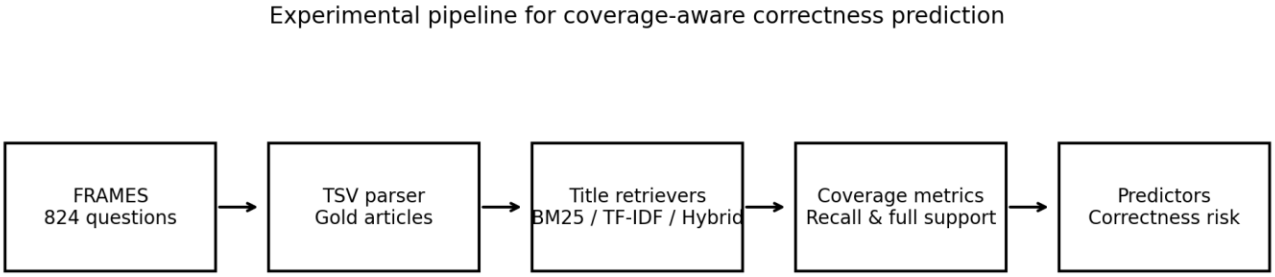

Figure 1. Experimental pipeline for coverage-aware correctness prediction.

Dataset Parsing and Evidence Structure

The parser reads wikipedia_link_1 through wikipedia_link_11+ and also scans the wiki_links field. Cells in wikipedia_link_11+ may contain several comma-separated URLs; these are expanded before deduplication. Across 824 questions, parsing yields 2,674 article mentions and 2,517 unique normalized titles. The mean evidence-set size is 3.25 articles, the median is three, and the observed range is two to 23. Table 1 summarizes the evaluation population.

Table 1. Parsed FRAMES dataset summary used in the experiment.

Dataset property	Value
Questions evaluated	824
Parsed gold article mentions	2,674
Unique parsed Wikipedia article titles	2,517
Mean articles per question	3.25
Median articles per question	3
Minimum / maximum parsed articles	2 / 23
Evaluation split	test.tsv
Random seed	2024

FRAMES assigns five reasoning categories, and a question can carry more than one label. Multiple constraints is the most common category, followed by numerical, temporal, tabular, and post-processing reasoning. Because the categories are multi-label, their counts sum to more than 824. Table 2 reports the distribution used for stratified analysis and prediction features.

Table 2. Reasoning-type distribution; categories are multi-label.

Reasoning type	Questions	Share of 824
Multiple constraints	549	66.6%
Numerical reasoning	293	35.6%
Post processing	107	13.0%
Tabular reasoning	236	28.6%
Temporal reasoning	278	33.7%

The parsed evidence sets are concentrated at two to four articles, but a small tail contains substantially more links. Table 3 groups questions by parsed article count. Figure 2 complements the binned table with the detailed evidence-count distribution and makes the long tail visible.

Table 3. Parsed gold article-count distribution after expanding wikipedia_link_11+.

Parsed article-count bin	Questions	Share
2	314	38.1%
3	286	34.7%
4	132	16.0%
5-6	55	6.7%
7-10	26	3.2%
11+	11	1.3%

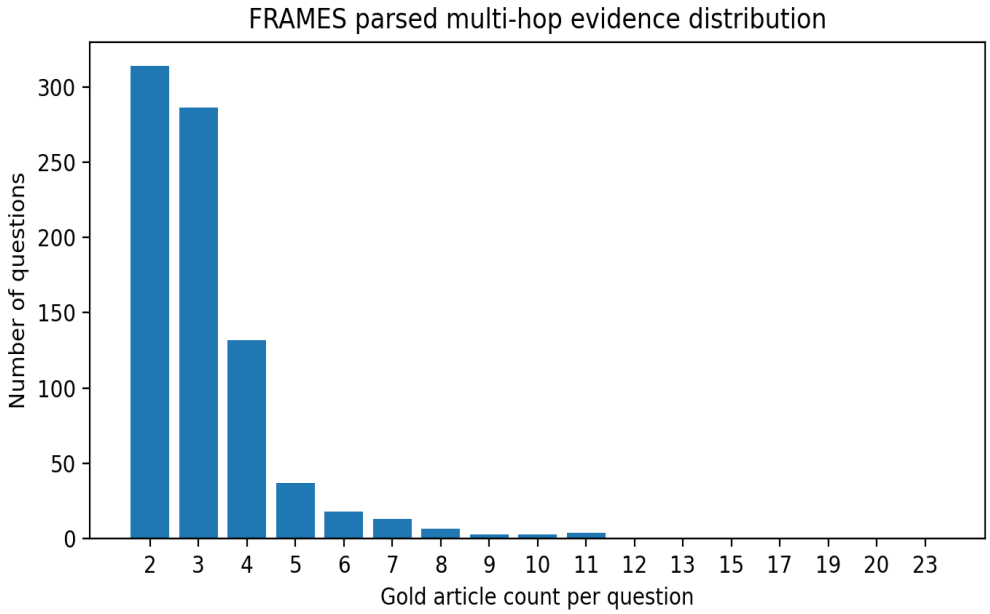

Figure 2. Parsed multi-hop evidence distribution in FRAMES.

Article Normalization and Retrieval Corpus

Each Wikipedia link is reduced to an article-level identity. The parser removes the scheme and domain, decodes percent escapes, drops fragment identifiers, replaces spaces with underscores, and deduplicates case-insensitively within each question. Section anchors are therefore treated as references to the same article. This choice follows the benchmark's article-level annotations and avoids counting multiple sections of one page as separate retrieval targets.

The candidate corpus is the union of all normalized gold titles in the test file. For tokenization, underscores are converted to spaces and punctuation is normalized. Article bodies, redirects, anchor text, and live Wikipedia content are not used. The design creates a closed-world document-discovery task in which every candidate is an annotated FRAMES article identity and the ranking depends only on the question and title representation.

Retrieval Models

Four title retrievers are evaluated. BM25-title applies Okapi BM25 with k1=1.5 and b=0.75 to tokenized titles [9]. TF-IDF-word-title uses word unigrams and bigrams with cosine similarity [8]. TF-IDF-char-title uses character n-grams of length three to five, which can recover punctuation, inflectional, and partial-string matches that word tokenization misses. Hybrid-title combines normalized BM25, word TF-IDF, and character TF-IDF scores with fixed weights of 0.45, 0.35, and 0.20. The fixed weights provide a deterministic comparison rather than a tuned ensemble.

The retrieval models rank all 2,517 candidate titles for each question. A title can receive a high score through an exact entity mention, a rare shared token, a phrase match, or character-level similarity. The models do not perform reasoning or query decomposition; their purpose is to reveal how much of the annotated document set can be recovered from the surface form of the original question.

Coverage and Strict Correctness

For question q, let Gq denote the parsed gold article set and Rq,k the top-k retrieved set. Coverage is |Gq intersection Rq,k| / |Gq|. Precision@k is |Gq intersection Rq,k| / k, and mean reciprocal rank uses the rank of the first retrieved gold article. Strict coverage-gated correctness is y(q,k)=1 when Gq is a subset of Rq,k and zero otherwise. Equivalently, y(q,k)=1 when coverage equals one.

The strict label represents a complete-evidence retrieval policy. A positive label means retrieval has not withheld any annotated supporting article from the downstream reader. A negative label identifies an evidence deficit, even when one or more relevant titles appear near the top. This distinction is central to the analysis because multi-hop readiness depends on the full set, whereas first-hit metrics depend on only one item.

Prediction Features and Models

The primary prediction target is strict correctness at k=10 for Hybrid-title. The operational feature set contains 18 variables available before generation: question-token count, character length, number of digits, a temporal-phrase flag, number of reasoning labels, five reasoning-type indicators, top-1 score, the top-1-to-top-2 score gap, top-10 score mean, top-10 minimum, top-10 score entropy, number of nonzero top-10 scores, and the mean and maximum Jaccard token overlap between the question and the retrieved titles. A secondary 19-feature setting adds parsed gold hop count.

Score entropy is calculated after nonnegative top-10 scores are normalized to sum to one. Low entropy indicates a concentrated ranking; high entropy indicates that score mass is spread across many titles. The top-two gap measures first-result dominance, while overlap summarizes how directly the ranking reflects prompt vocabulary. The predictor never receives the gold hit count, coverage value, or strict label as an input.

Logistic regression, random forest, and gradient boosting are implemented with scikit-learn [24]. Logistic regression standardizes continuous features and uses class-balanced weights. The random forest uses 100 trees, maximum depth seven, minimum leaf size five, and balanced class weights. Gradient boosting uses 100 estimators, learning rate 0.06, and maximum depth two. Hyperparameters and random seed are held fixed across comparisons.

Evaluation Protocol

Retrieval is evaluated at k=4, 8, 10, 15, and 25 to measure the trade-off between compact context and evidence recall. Prediction uses stratified five-fold cross-validation with random seed 2024. For Hybrid-title at k=10, 185 of 824 questions are positive and 639 are negative. The folds preserve this class ratio. Reported prediction metrics are ROC-AUC, PR-AUC, Brier score, accuracy, and F1. The same folds are used for model and feature comparisons so that differences reflect the feature or estimator rather than a different partition.

The operational setting represents information that a deployed controller can compute from the question and ranked list. Reasoning labels are available in FRAMES and are treated as supplied query metadata. The hop-count setting is reported separately because gold evidence-set size is not normally known at deployment; it estimates the upper value of an accurate hop-count signal rather than an operational baseline.

Results and Discussion

Retrieval Coverage

Table 4 reports mean gold-article coverage for every retriever and cutoff. Hybrid-title has the highest mean coverage at each evaluated k, although its margin over BM25-title is small. At k=10, the hybrid reaches 0.547, compared with 0.546 for BM25, 0.521 for character TF-IDF, and 0.468 for word TF-IDF. At k=25, hybrid

coverage reaches 0.590. Figure 3 shows that all methods improve monotonically as the context budget grows, but the curves flatten well below complete coverage.

Table 4. Mean retrieval coverage by method and top-k.

Retriever	Coverage@4	Coverage@8	Coverage@10	Coverage@15	Coverage@25
BM25-title	0.484	0.533	0.546	0.565	0.582
Hybrid-title	0.488	0.538	0.547	0.571	0.590
TF-IDF-char-title	0.453	0.506	0.521	0.548	0.581
TF-IDF-word-title	0.403	0.454	0.468	0.484	0.513

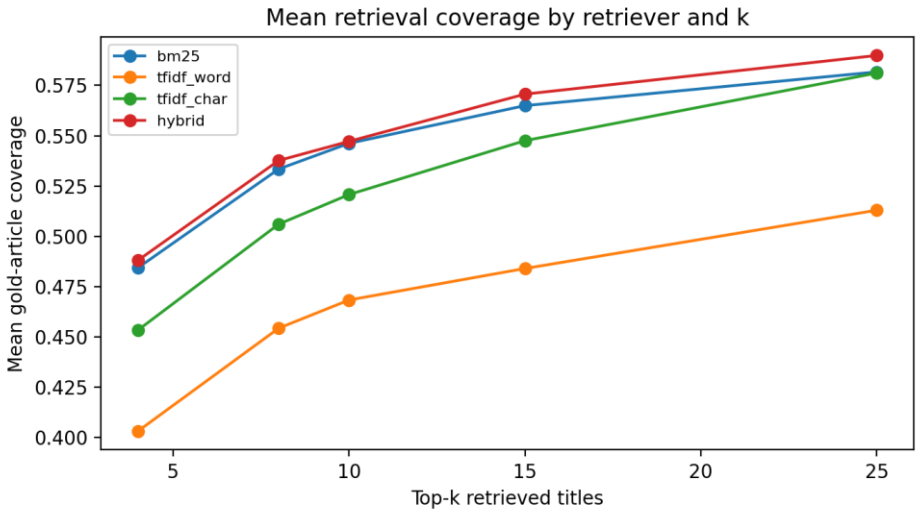

Figure 3. Mean retrieval coverage by retriever and top-k.

The ranked outputs also show why coverage adds information beyond first-hit success. BM25-title and Hybrid-title place the first relevant article near the top for many questions, producing first-relevant mean reciprocal rank above 0.82. Hybrid precision@10 is only 0.159, however, and average coverage remains near one half. The retriever often finds a strong anchor page while allocating the remaining slots to related but non-gold titles or missing another evidence branch entirely.

Strict Coverage-Gated Correctness

Table 5 converts partial coverage into complete-evidence success. Hybrid-title reaches strict correctness of 0.225 at k=10 and 0.273 at k=25. BM25 is tied at k=10 and slightly lower at k=25. Character TF-IDF becomes competitive as k grows, while word TF-IDF remains the weakest method. Figure 4 shows that strict correctness rises much more slowly than mean coverage because every missing article keeps the label at zero.

Table 5. Strict coverage-gated correctness by method and top-k.

Retriever	Strict@4	Strict@8	Strict@10	Strict@15	Strict@25
BM25-title	0.174	0.218	0.225	0.243	0.263
Hybrid-title	0.174	0.217	0.225	0.250	0.273
TF-IDF-char-title	0.155	0.200	0.218	0.239	0.265
TF-IDF-word-title	0.131	0.167	0.178	0.191	0.216

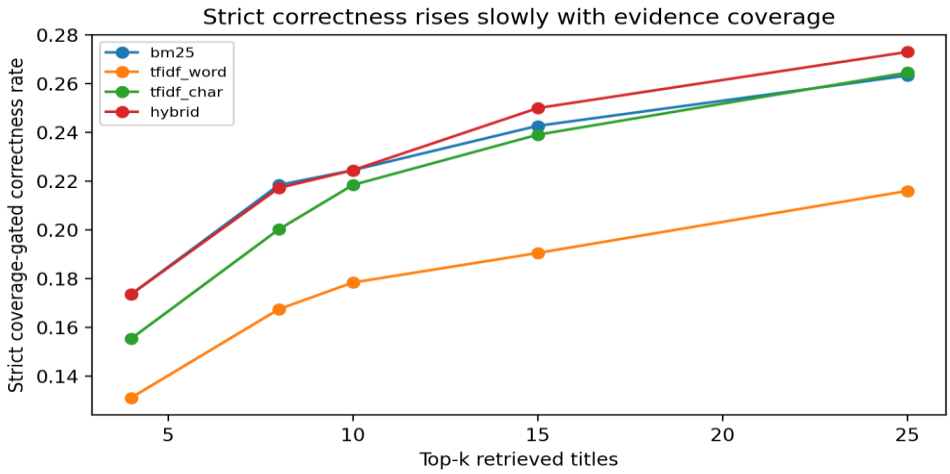

Figure 4. Strict coverage-gated correctness by retriever and top-k.

Expanding the hybrid context from ten to 25 titles increases mean coverage by 0.043 but strict correctness by only 0.048. The gain is real, yet most questions still lack at least one annotated article. A flat increase in k therefore cannot replace retrieval strategies that identify distinct entities, constraints, or intermediate targets. It also introduces more distractors and consumes a larger context budget.

Evidence-Set Size and Reasoning Type

The hop-count analysis in Table 6 explains much of the strict-correctness gap. At k=10, two-article questions reach coverage of 0.683 and strict correctness of 0.449. Three-article questions fall to 0.519 coverage and 0.112 strict correctness, while four-article questions fall to 0.420 and 0.053. The eleven-plus bin reaches only 0.188 coverage and no strict positives. Figure 5 visualizes the decline at k=10 and k=25.

Table 6. Hybrid-title coverage and strict correctness by parsed article-count bin.

Hop bin	Questions	Cov@10	Strict@10	Cov@25	Strict@25
2	314	0.683	0.449	0.720	0.510
3	286	0.519	0.112	0.561	0.154
4	132	0.420	0.053	0.479	0.098
5-6	55	0.410	0.055	0.446	0.073
7-10	26	0.306	0.077	0.349	0.154
11+	11	0.188	0.000	0.261	0.000

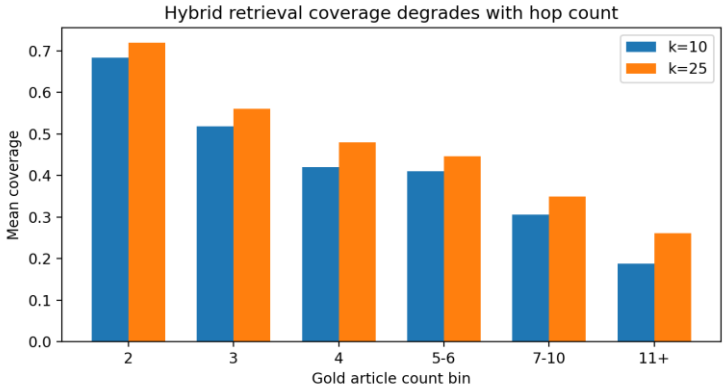

Figure 5. Hybrid-title coverage by parsed article-count bin.

The small high-hop bins should be interpreted cautiously, and their strict rates are not perfectly monotonic because they contain few questions. The broad pattern is nevertheless clear: each additional annotated article creates

another opportunity for omission. Widening k helps every coverage bin, but even k=25 recovers only 0.261 of the gold set on average for the eleven-plus group.

Reasoning type provides a second view of retrieval risk. Table 7 shows that numerical reasoning has the highest Hybrid-title coverage and strict correctness at k=10, whereas multiple-constraint questions are the hardest. The latter category covers 549 questions and often requires the retriever to coordinate several entities and filters. A ranking may locate the prominent entity named or described in the question while missing a comparison page, list, or table needed to apply the remaining constraint.

Table 7. Hybrid-title k=10 results by reasoning type.

Reasoning type	Questions	Hybrid Cov@10	Hybrid Strict@10
Multiple constraints	549	0.481	0.129
Numerical reasoning	293	0.610	0.317
Post processing	107	0.511	0.178
Tabular reasoning	236	0.543	0.229
Temporal reasoning	278	0.554	0.201

Predicting Strict Correctness

Table 8 reports five-fold prediction performance at k=10. On Hybrid-title with operational features, logistic regression reaches ROC-AUC 0.796, PR-AUC 0.590, Brier score 0.178, accuracy 0.733, and F1 0.534. Random forest and gradient boosting have similar ranking performance, while gradient boosting achieves the lowest operational Brier score at 0.142. Figure 6 compares the operational ROC curves and shows consistent separation between complete- and incomplete-evidence cases.

Table 8. Five-fold prediction of strict correctness at k=10.

Setting	Predictor	N	Pos. rate	ROC-AUC	PR-AUC	Brier	Accuracy	F1
Hybrid / operational	Logistic regression	824	0.225	0.796	0.590	0.178	0.733	0.534
Hybrid / + hop count	Logistic regression	824	0.225	0.856	0.689	0.150	0.778	0.605
Hybrid / operational	Random forest	824	0.225	0.768	0.516	0.161	0.780	0.517
Hybrid / + hop count	Random forest	824	0.225	0.864	0.686	0.128	0.807	0.599
Hybrid / operational	Gradient boosting	824	0.225	0.774	0.518	0.142	0.794	0.370
Hybrid / + hop count	Gradient boosting	824	0.225	0.860	0.673	0.116	0.845	0.556
All retrievers / operational	Logistic regression	3,296	0.211	0.807	0.580	0.175	0.737	0.529
All retrievers / operational	Random forest	3,296	0.211	0.851	0.650	0.148	0.801	0.596
All retrievers / operational	Gradient boosting	3,296	0.211	0.836	0.619	0.122	0.826	0.407

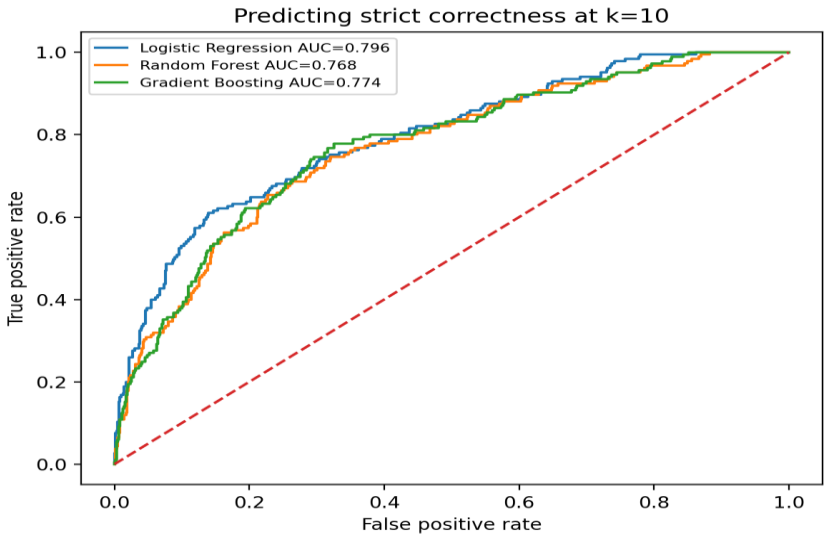

Figure 6. ROC curves for operational Hybrid-title correctness prediction at k=10.

Adding hop-count metadata improves every model. Random forest reaches the highest ROC-AUC, 0.864, while logistic regression reaches ROC-AUC 0.856 and F1 0.605. The improvement confirms that evidence-set size is a major determinant of strict success. It also defines an opportunity for deployment: a system that can estimate required hop count from the prompt may recover part of this gain without access to gold annotations.

Pooling all four retrievers increases the sample to 3,296 query-retriever pairs and allows method indicators to capture retrieval-specific behavior. Random forest reaches ROC-AUC 0.851 and PR-AUC 0.650 in this setting. The pooled result suggests that coverage risk is not confined to one score function, although calibration and thresholds would need to be recalibrated when the corpus or retriever changes.

Feature Ablation, Importance, and Calibration

Table 9 separates the information contributed by prompt, reasoning, score, and overlap features. Prompt-only features reach ROC-AUC 0.675, reasoning labels reach 0.708, and score-distribution features reach 0.740. Overlap alone is weaker at 0.609. Combining scores and overlap yields 0.754, while the full operational set reaches 0.796. The result shows that no single signal is sufficient: the strongest prediction combines question structure with ranking geometry and lexical alignment.

Table 9. Logistic-regression feature ablation on Hybrid-title strict correctness at k=10.

Feature set	Features	ROC-AUC	PR-AUC	Brier	Accuracy	F1
Prompt only	5	0.675	0.410	0.227	0.610	0.422
Reasoning types only	5	0.708	0.412	0.210	0.716	0.491
Score distribution only	6	0.740	0.492	0.202	0.697	0.488
Overlap only	2	0.609	0.296	0.242	0.624	0.392
Scores + overlap	8	0.754	0.509	0.197	0.701	0.494
All operational	18	0.796	0.590	0.178	0.733	0.534
All operational + hop count	19	0.856	0.689	0.150	0.778	0.605

Gradient-boosting importance provides a complementary nonlinear view. Table 10 lists the ten most important operational features. The multiple-constraints indicator is highest, followed by the top-two score gap, top-10 score entropy, top-10 minimum score, and mean title overlap. Figure 7 displays the broader importance profile used to interpret the model.

Table 10. Top gradient-boosting feature importances for Hybrid-title operational prediction.

Feature	Importance
rt_Multiple constraints	0.262
top2_gap	0.209
top10_score_entropy	0.131
top10_min_score	0.108
top10_overlap_mean	0.075
top1_score	0.038
n_prompt_chars	0.036
n_digits	0.030
rt_Numerical reasoning	0.025
rt_Temporal reasoning	0.022

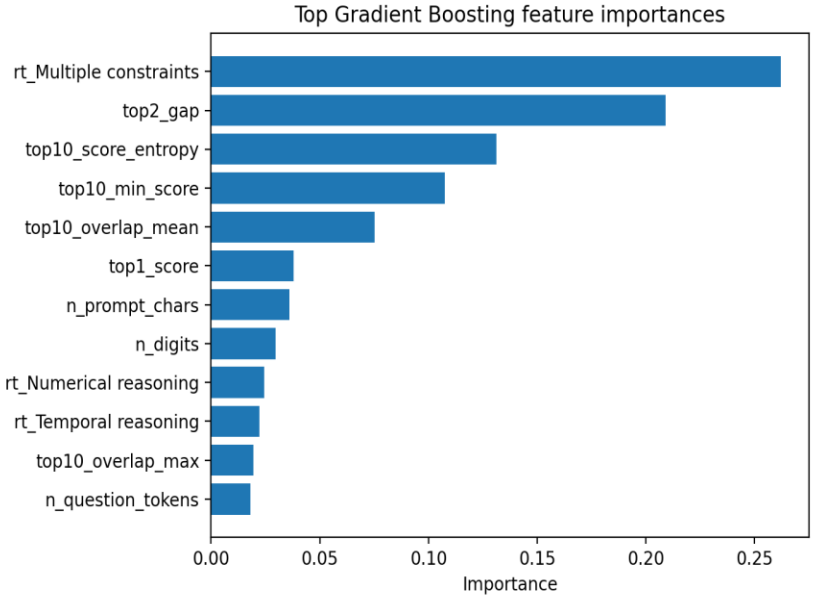

Figure 7. Top feature importances from the gradient-boosting predictor.

The importance pattern is consistent with the coverage mechanism. A large top-two gap can indicate that the retriever found one obvious anchor while the rest of the evidence remains weak. High entropy can instead indicate a diffuse ranking with no clear evidence structure. The multiple-constraints indicator marks questions whose gold set is likely to span several semantic branches. Overlap and tail-score features help distinguish a broadly relevant top-10 list from one dominated by a single lexical match.

Probability quality matters because the predictor is intended to control retrieval actions. Figure 8 shows the Hybrid-title gradient-boosting calibration curve. Its operational Brier score of 0.142 is lower than that of logistic regression and random forest, even though its F1 at a fixed 0.5 threshold is lower. The difference illustrates why a controller should evaluate both discrimination and calibration: ranking metrics determine whether risky cases can be ordered, whereas calibrated probabilities support stable action thresholds.

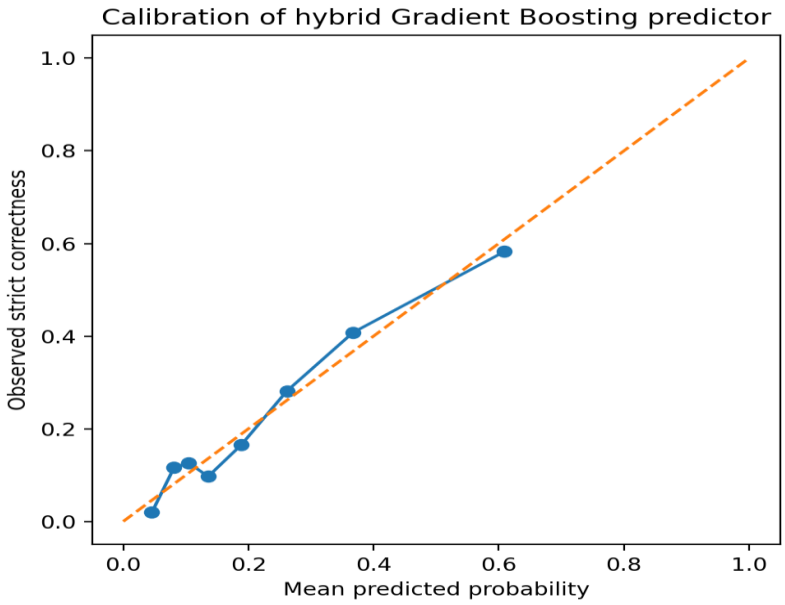

Figure 8. Calibration curve for the Hybrid-title gradient-boosting predictor.

Implications for Multi-Hop RAG

The experiments support three conclusions. First, first-hit retrieval quality does not guarantee multi-hop readiness. A system can rank one gold title early and still fail because another article is absent. Second, increasing k provides diminishing returns for complete evidence, particularly as the annotated article set grows. Third, the risk of incompleteness is predictable from inexpensive signals available before generation.

These findings motivate a coverage-aware controller. After an initial ranking, the controller can estimate the probability that the context contains the complete evidence set. High predicted support can permit direct generation with normal citation checks. Intermediate support can trigger targeted query decomposition or iterative retrieval for missing entities and constraints. Low predicted support can trigger a deeper search procedure, a request for clarification, or abstention. The action thresholds should be calibrated for the costs and retrieval distribution of the deployment domain.

The results also clarify the role of iterative multi-hop methods. Their benefit is not only that they produce longer reasoning traces; they can allocate retrieval steps to evidence branches that a single flat ranking misses. The sharp decline from two-article to three- and four-article questions gives a concrete target for such methods: improve set completion rather than repeatedly returning variants of the same anchor article.

Reporting coverage and answer correctness together can improve RAG diagnosis. A reader that fails despite full evidence has a reasoning or synthesis problem. A reader that succeeds despite incomplete annotated coverage may be using parametric knowledge, unannotated evidence, or redundancy in the gold set. A reader that fails with incomplete evidence has a retrieval bottleneck. Separating these cases makes system improvement more precise than an aggregate end-to-end accuracy score alone.

Limitations

The first limitation is the title-level corpus. FRAMES provides article links but not article body text in the test file used here. A production retriever can exploit summaries, redirects, anchor text, tables, aliases, and passage content that are absent from normalized titles. The reported coverage values should therefore be interpreted as closed-world title-discovery results rather than as an estimate of full-web or passage-level retrieval performance.

The second limitation is the strict correctness definition. Some questions may be answerable from a subset of the annotated pages because the evidence is redundant, while other questions may require calculations or comparisons that remain difficult even after every page is retrieved. Complete article coverage is consequently neither a

universal necessary condition nor a sufficient condition for every reader. It is a conservative, transparent indicator of whether the retriever supplied the benchmark's full annotated evidence set.

The third limitation concerns annotation and parsing variability. The wikipedia_link_11+ field can contain several URLs in one cell, which produces an observed maximum of 23 parsed titles. High-count bins are small: only 11 questions fall in the eleven-plus group. Their exact strict rates should be viewed as diagnostics, while the larger two-, three-, and four-article groups provide more stable evidence for the overall hop-count pattern.

The fourth limitation is that the predictors are trained and tested within one benchmark and retrieval distribution. They may learn FRAMES-specific regularities in question wording, reasoning labels, and title structure. A monitor deployed with dense retrieval, passage retrieval, a different corpus, or another language would require new calibration data and an assessment of feature stability.

Finally, the candidate corpus is constructed from the union of gold-linked article identities in the evaluation file. This design creates a controlled ranking problem but does not model open-world indexing, corpus freshness, or the presence of millions of unrelated pages. Future work should repeat the analysis with archived article text, a larger fixed Wikipedia snapshot, iterative retrievers, and end-to-end readers so that retrieval coverage can be connected directly to answer accuracy and citation faithfulness.

Conclusion

This study evaluated retrieval coverage and coverage-risk prediction on all 824 FRAMES questions. A closed-world title corpus was constructed from parsed Wikipedia links, and BM25, word TF-IDF, character TF-IDF, and a fixed hybrid were compared at several retrieval depths. Hybrid-title achieved the strongest mean coverage, reaching 0.547 at k=10 and 0.590 at k=25. Strict coverage-gated correctness remained substantially lower, at 0.225 and 0.273, because one missing article is enough to break a complete-evidence chain.

Evidence-set size is the clearest structural risk factor. Two-article questions are often recoverable with a flat ranking, while three-article, four-article, and higher-hop questions experience a steep decline in complete coverage. Multiple-constraint questions are particularly difficult because their evidence tends to span several entities, filters, or tables. Increasing k helps, but it does not remove the need for branch-aware retrieval.

Coverage failure is predictable before generation: ROC-AUC is 0.796 with operational features and 0.856 with hop-count metadata. This estimate can guide answering, further retrieval, decomposition, or abstention.

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," arXiv:2409.12941, 2024.

[2] Google, "google/frames-benchmark," Hugging Face Datasets, 2024.

[3] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.

[4] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. ICML, 2020.

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

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

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

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.

[9] 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.

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

[11] 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.

[12] J. Welbl, P. Stenetorp, and S. Riedel, "Constructing Datasets for Multi-hop Reading Comprehension Across Documents," Transactions of the Association for Computational Linguistics, vol. 6, pp. 287-302, 2018.

[13] W. Chen et al., "HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data," in Proc. EMNLP, pp. 1026-1036, 2020.

[14] 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.

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

[16] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. EACL, pp. 874-880, 2021.

[17] G. Izacard et al., "Few-shot Learning with Retrieval Augmented Language Models," Journal of Machine Learning Research, vol. 24, no. 251, pp. 1-43, 2023.

[18] 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

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

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

[21] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering," in Proc. EMNLP, pp. 2381-2391, 2018.

[22] K. Cobbe et al., "Training Verifiers to Solve Math Word Problems," arXiv:2110.14168, 2021.

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

[24] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[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] 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.

[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] 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.

[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] 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.

[33] 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.

[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] 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.

[36] J. Chen, X. Sun, Q. Wu, and M. Jackson, "Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID," JACS, vol. 4, no. 4, pp. 61-79, Apr. 2024, doi: 10.69987/JACS.2024.40406.

[37] 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.

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