JOURNAL OF GLOBAL ENGINEERING REVIEW ISSN: 3071-4656

Nugget-Level Retrieval Coverage Prediction for Complete and Faithful RAG Answers

David Kim
Computer Science, University of Southern California, Los Angeles, CA, USA
dkim_usc2025@yahoo.com

Abstract
Retrieval-augmented generation (RAG) depends on retrieved passages to provide the evidence from which an answer is composed. Conventional retrieval metrics reward relevance, yet a highly relevant top-k list can repeatedly surface the same fact while omitting other information units required for a complete response. This study investigates nugget-level retrieval coverage prediction on CoverageBench, which comprises seven retrieval collections, 334 topics, and six retrieval configurations. Using the benchmark's aggregate nDCG@20, alpha-nDCG@20, and Subtopic Recall at rank 20 (StRecall@20) measurements, we introduce NLCP-Ridge, a deterministic ridge-regression model that predicts coverage from relevance, diversity-aware relevance, and reranker identity. Leave-one-dataset-out validation over 42 held-out system-dataset cases yields MAE = 0.0464, RMSE = 0.0687, R2 = 0.9090, Pearson r = 0.9628, and Spearman rho = 0.9713. Relative to an nDCG-only predictor, NLCP-Ridge reduces MAE by 58.6%; relative to an alpha-nDCG-only predictor, it reduces MAE by 18.0%. The retrieval comparisons further show that the highest-nDCG configuration is not always the highest-coverage configuration: CAsT and CRUX-MultiNews select different winners under StRecall@20. The findings show that a compact coverage estimate can complement relevance evaluation and provide an evidence-sufficiency signal before answer generation.

Keywords: *Retrieval-augmented generation; information coverage; nugget-level evaluation; Subtopic Recall; alpha-nDCG; retrieval prediction; faithful generation; CoverageBench.*

1. Introduction

Retrieval-augmented generation (RAG) combines information retrieval with language generation so that answers can draw on evidence outside model parameters. The original formulation conditions generation on retrieved passages [1], while dense passage retrieval [2], passage-fusion architectures [3], broader RAG taxonomies [4], and self-reflective retrieval-generation loops [5] have expanded the design space. Across these variants, retrieval remains the principal gatekeeper: the generator can only ground a claim in evidence that has reached its context window.

This dependency creates a distinction between relevance and evidential completeness. A ranked list may contain highly relevant documents yet devote most of its limited context budget to one aspect of a multifaceted question. Standard RAG evaluation frameworks assess context relevance, answer relevance, and faithfulness [6], [7], while atomic-fact evaluation asks whether generated claims are supported [8]. These perspectives are essential, but they do not by themselves determine whether the retrieved set contains every major fact, caveat, comparison, or subgroup needed for a complete answer.

Information-retrieval research has long recognized that ranked results should not be treated as independent items when redundancy matters. Maximal marginal relevance balances relevance and novelty [9]; diversity-aware evaluation discounts repeated subtopic coverage [10]-[13]; and nugget-based question answering and pyramid-based summarization evaluate content at the level of atomic information units [14], [15]. For RAG, this set-based

view is especially important because repeated evidence consumes tokens without increasing the range of answerable facts.

Coverage also clarifies the relationship between faithfulness and completeness. An answer can be faithful but incomplete when it accurately summarizes a narrow evidence set. Conversely, an apparently complete answer becomes unfaithful when it supplies missing facts without support. A retrieval-stage coverage estimate therefore addresses a failure mode that claim verification alone cannot resolve: whether the evidence set is broad enough to support the full scope of the requested answer.

The central research question is whether nugget coverage can be predicted from routinely reported retrieval signals strongly enough to support operational decisions. A useful predictor should transfer across collections, remain interpretable, and estimate coverage before an answer is generated. A low predicted StRecall can then trigger query decomposition, diversified reranking, an expanded retrieval budget, or an explicit evidence-sufficiency warning.

This study makes three contributions. First, it provides a complete comparison of six retrieval configurations across all seven CoverageBench datasets. Second, it introduces NLCP-Ridge, a compact predictor trained and evaluated under leave-one-dataset-out validation. Third, it connects prediction accuracy with retrieval-system choice through dataset-level error analysis, metric correlations, winner disagreements, and coverage-gap diagnostics. Together, these analyses show when relevance is a reliable proxy for coverage and when it is not.

2. Literature Review

2.1 RAG architectures and evaluation

RAG research initially focused on coupling a retriever with a generator [1], improving first-stage retrieval [2], and integrating multiple retrieved passages during generation [3]. Subsequent surveys organized the field around retrieval, augmentation, and generation components [4], while Self-RAG introduced learned decisions about when to retrieve and how to critique retrieved evidence [5]. These developments improved grounding, but they also made evaluation more layered: retrieval quality, context usefulness, answer support, and answer completeness can fail independently.

Automated evaluation frameworks reflect this layered structure. RAGAS evaluates dimensions such as context relevance and faithfulness without requiring a reference answer [6], and ARES trains lightweight judges for context relevance, answer faithfulness, and answer relevance [7]. FActScore decomposes long-form text into atomic facts and measures the fraction supported by a reliable source [8]. These methods are well suited to diagnosing unsupported content; coverage evaluation adds the complementary question of whether the evidence and answer address the full set of required information units.

2.2 Retrieval effectiveness, reranking, and diversification

The retrieval models represented in modern RAG systems span lexical, dense, sparse-neural, and reranking approaches. BM25 remains a strong lexical baseline [16]. Dense passage retrieval learns a dual-encoder representation for open-domain retrieval [2], Sentence-BERT provides sentence-level embeddings [21], ColBERT introduces contextualized late interaction [19], and SPLADE combines learned expansion with sparse retrieval [20]. Cross-encoder and contextualized rerankers such as BERT passage reranking [22] and CEDR [23] improve ordering after candidate generation, while Pyserini supports controlled comparison of sparse and dense pipelines [24].

Traditional effectiveness measures remain indispensable. nDCG evaluates graded relevance with rank discounting [17], and rank-biased precision models user persistence down a result list [18]. Their limitation for coverage is structural rather than methodological: they primarily reward the relevance of individual results. MMR explicitly penalizes redundancy [9], alpha-nDCG rewards new subtopic gain [10], TREC Web Track evaluations operationalized diversity across query intents [11], intent-aware measures formalized this objective [12], and cascade measures examined alternative user models for novelty [13]. Nugget evaluation in the TREC Question

Answering Track [14] and the pyramid method for summarization [15] further shifted evaluation from documents or sentences toward the atomic information units that a complete response should contain.

2.3 Coverage-centered evaluation and retrieval

Recent RAG research has made coverage an explicit target. Xie et al. evaluate open-ended answers through sub-question coverage and show that retrieval of core sub-questions can improve response quality [25]. CoverageBench unifies seven collections with nugget or subtopic annotations and reports six common retrieval configurations under relevance and coverage metrics [26]. The Great Nugget Recall develops an automated nugget-extraction and assignment framework for long-form RAG evaluation [27], reducing the annotation burden associated with atomic-fact scoring.

Coverage-aware methods also move beyond measurement. LANCER generates proxy sub-questions, predicts document answerability, and reranks for nugget coverage [28]. CoveR trains a dense retriever with coverage-sensitive objectives derived from sub-question answerability [29]. These studies demonstrate that coverage can be measured and directly optimized. The remaining operational gap is to estimate expected coverage when complete nugget judgments are unavailable for a new run or collection. The present study addresses that gap with a small transfer-oriented predictor built from aggregate retrieval signals.

2.4 Evidence-grounded RAG in applied settings

Applied RAG studies make the cost of under-coverage concrete. In financial analysis, evidence-grounded risk memos over SEC filings combine retrieval with numeric verification [30], related work targets numerical hallucination in corporate risk reporting [31], and tokenized receivables analysis requires retrieval that respects disclosure and capital-market constraints [32]. Numerical-reasoning guardrails for quantitative research assistants similarly emphasize that a plausible answer is not sufficient when a material figure or filing context is absent [38]. Across these applications, retrieval must cover complementary qualitative and quantitative evidence rather than merely return topically similar passages.

Operational domains pose a parallel challenge. Cloud-native DevOps question answering uses evidence-grounded retrieval over private operations documents [33], and OpsLLM combines bilingual retrieval with incident triage and alert summarization [34]. Long-document RAG for contractual and insurance clauses must preserve dispersed obligations and exceptions [36], while retrieval-summary integration for code intelligence separates finding relevant material from explaining it [37]. A budgeted multi-hop retrieval agent extends this logic to compositional questions in which evidence must be collected across multiple hops [39]. These systems motivate a pre-generation estimate of whether the retrieved context is sufficiently broad for synthesis.

A third line of work focuses on verification and human oversight. Evidence-chain RAG links hallucination detection with source attribution and provenance explanation [35]. Lightweight evidence-consistency checks and small-model detection can block unsupported enterprise responses [40], and narrative-aware claim-verification agents rank evidence for scientific claims [41]. Trust-calibrated multilingual RAG adds selective abstention and interface-level communication for humanitarian information access [42]. Such safeguards are strongest when paired with a retrieval diagnostic that identifies missing evidence before generation rather than only detecting unsupported claims afterward.

2.5 Study positioning

The literature therefore provides four complementary components: strong relevance-oriented retrieval, diversity-aware ranking, nugget-level coverage evaluation, and evidence-grounded generation controls. NLCP-Ridge is positioned between evaluation and control. It does not replace StRecall annotations, rerank documents directly, or verify generated claims. Instead, it learns the empirical relationship between relevance, diversity-aware relevance, reranking behavior, and held-out coverage so that an RAG pipeline can estimate evidence breadth before committing to generation.

3. Method

3.1 Benchmark and retrieval configurations

The analysis uses the public aggregate result matrix from CoverageBench [26]. The benchmark contains 334 topics distributed across NeuCLIR, RAG, Fair Ranking, CAsT, CRUX-MultiNews, CRUX-DUC04, and RAGTIME. Six retrieval configurations are evaluated on every dataset: BM25, BM25 followed by Qwen3 reranking, BM25 followed by Rank1 reranking, Qwen3-8B, Qwen3-8B followed by Qwen3 reranking, and Qwen3-8B followed by Rank1 reranking. The full factorial design yields 42 system-dataset observations.

Table 1 summarizes the collection-level descriptors. Collection scale varies from 565,015 shared CRUX documents to more than 113 million RAG passages, while the average number of target nuggets ranges from 6.1 for CAsT to 29.7 for Fair Ranking. This variation creates a demanding transfer setting: each held-out dataset differs in corpus scale, passage length, query count, and nugget density.

Table 1. Dataset statistics in the seven-dataset CoverageBench evaluation.

Dataset	Docs/passages	Avg words	Queries	Avg nuggets	Med nuggets	Min	Max
NeuCLIR	10,038,768	348.0	19	14.9	15.0	10	25
RAG	113,520,750	166.7	56	13.9	14.0	6	20
Fair Rank	6,475,537	479.2	50	29.7	26.0	3	62
CAsT	38,429,852	59.6	25	6.1	6.0	3	12
CRUX-MultiNews	565,015	88.9	100	14.2	14.0	12	15
CRUX-DUC04	565,015	88.9	50	7.8	8.0	1	10
RAGTIME	4,000,380	404.0	34	15.6	16.5	8	20

3.2 Metrics and coverage target

Three rank-20 metrics define the modeling problem. nDCG@20 measures graded relevance with position discounting [17]. Alpha-nDCG@20 additionally discounts repeated coverage of subtopics and therefore rewards early novelty [10], [12]. StRecall@20 is the prediction target and measures the fraction of distinct target nuggets covered by at least one document in the top 20 [26]. For topic q, the metric can be written as StRecall@20(q) = |CoveredNuggets(q, 20)| / |N(q)|, where N(q) is the topic's target nugget set.

The three metrics answer different questions. nDCG asks whether highly ranked results are relevant; alpha-nDCG asks whether those results contribute new subtopic information; StRecall asks whether the retrieved set reaches the target information inventory. A list can therefore score well on relevance yet leave important nuggets uncovered, or obtain broad binary coverage while placing the most useful evidence relatively low in the ranking.

3.3 NLCP-Ridge

NLCP-Ridge predicts StRecall@20 from two numeric inputs, nDCG@20 and alpha-nDCG@20, plus a categorical reranker indicator with three levels: no reranker, Qwen3 reranking, or Rank1 reranking. Numeric features are standardized within each training fold, and reranker identity is one-hot encoded. The transformed features are fitted with ridge regression using alpha = 0.03. Ridge regression stabilizes coefficients in the presence of correlated predictors while retaining a transparent linear mapping [44].

The feature set is intentionally compact. Dataset descriptors are informative about collection scale but cannot describe how a retrieval system behaved on that collection. System identity alone cannot capture dataset difficulty. By combining observed relevance, observed diversity-aware relevance, and reranker family, NLCP-Ridge

estimates the remaining systematic relationship with coverage without fitting dataset-specific intercepts that would be unavailable for a new collection.

3.4 Validation protocol and comparison models

Generalization is evaluated with leave-one-dataset-out validation, an application of cross-validatory assessment in which an entire collection is excluded from training [43]. In each of seven folds, the model is trained on 36 observations from six datasets and evaluated on the six retrieval configurations of the held-out dataset. Every system-dataset pair appears exactly once in the test condition. Predictions are clipped to the valid interval from 0 to 1.

Performance is reported with mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R2), Pearson correlation, and Spearman rank correlation. Table 2 lists the comparison models. Mean-only provides a calibration floor; nDCG-only and alpha-nDCG-only isolate the two metric families; nDCG+alpha combines both metrics with their difference and ratio; dataset priors and system priors test whether collection metadata or architecture labels suffice without observed metric performance.

Table 2. Compared predictors and input signals.

Predictor	Inputs	Purpose
Mean-only	Training-fold mean StRecall@20	Calibration floor
nDCG-only	nDCG@20	Relevance-only estimate
alpha-nDCG-only	alpha-nDCG@20	Diversity-aware estimate
nDCG+alpha	nDCG@20, alpha-nDCG@20, gap, ratio	Metric-fusion baseline
Dataset priors	Collection scale, length, query, and nugget descriptors	Collection-level prior
System priors	Initial-ranker and reranker indicators	Architecture prior
NLCP-Ridge	nDCG@20, alpha-nDCG@20, reranker indicator	Proposed coverage predictor

3.5 Analysis workflow

Figure 1 summarizes the evaluation workflow. Each fold separates one dataset, derives metric signals and reranker identity for the training observations, applies fold-specific preprocessing, fits the ridge model, and predicts StRecall@20 for the held-out configurations. The analysis was implemented locally in Python with deterministic preprocessing and regression; no stochastic optimization is involved.

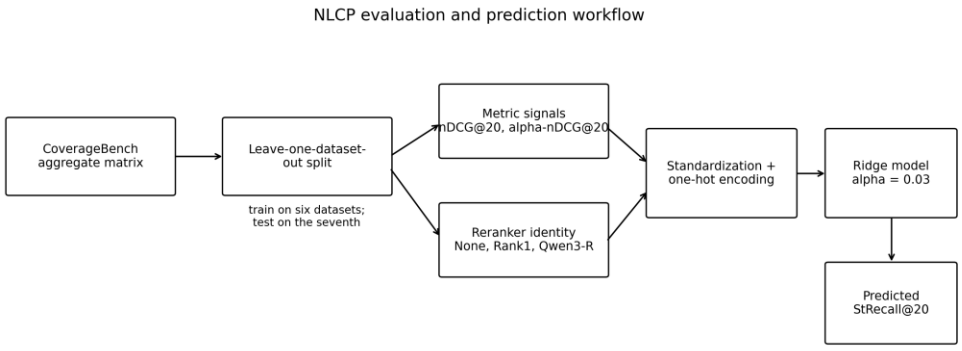

Figure 1. NLCP evaluation and prediction workflow.

4. Results and Discussion

4.1 Dataset coverage profiles

The benchmark spans markedly different coverage regimes. Figure 2 visualizes the average nugget count from Table 1. Fair Ranking has 29.7 nuggets per query on average, almost five times the 6.1 average for CAsT. A top-20 list can therefore approach full coverage on CAsT with relatively few distinct facets, whereas the same cutoff faces a much larger target inventory on Fair Ranking. RAGTIME, NeuCLIR, CRUX-MultiNews, and RAG occupy an intermediate range near 14 to 16 nuggets per query.

Collection size does not determine coverage difficulty by itself. RAG is the largest collection, yet it is one of the easiest held-out folds for NLCP-Ridge. Fair Ranking is substantially smaller but produces the largest prediction error. The contrast indicates that facet structure, lexical matching behavior, and nugget density matter more to cross-dataset transfer than document count alone.

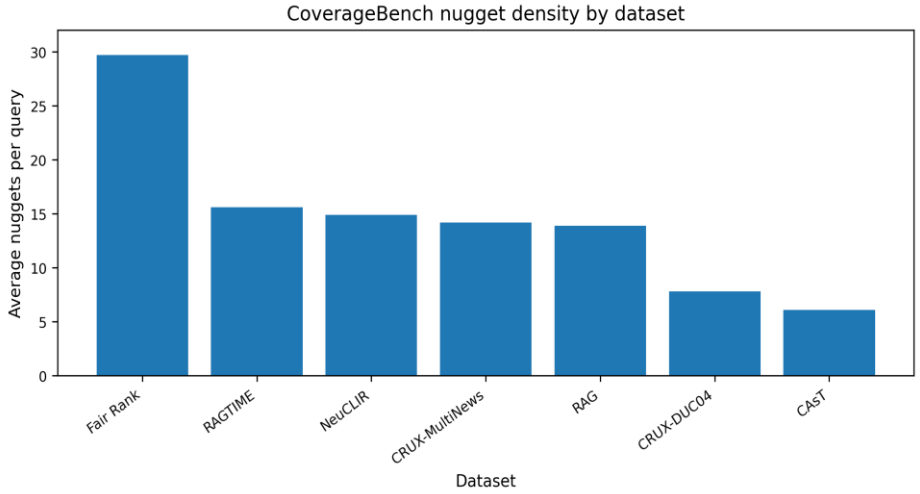

Figure 2. CoverageBench nugget density by dataset.

4.2 Relevance, diversity, and observed coverage

Table 3 reports nDCG@20 for all 42 observations. Qwen3-8B+Rank1 has the highest macro average at 0.676, followed by Qwen3-8B at 0.650 and Qwen3-8B+Qwen3-R at 0.632. BM25 has the lowest macro average at 0.431. Fair Ranking is the principal exception to the broader dense-retrieval advantage: BM25+Rank1 reaches the best nDCG@20 on that dataset at 0.161, while Qwen3-8B+Qwen3-R falls to 0.052. This pattern is consistent with the benchmark's characterization of Fair Ranking as a more syntactic, facet-explicit retrieval task [26].

Table 3. nDCG@20 for all datasets and retrieval configurations.

System	NeuCLIR	RAG	Fair Rank	CAsT	CRUX-MultiNews	CRUX-DUC04	RAGTIME	Macro avg
BM25	0.328	0.599	0.143	0.475	0.429	0.448	0.596	0.431
BM25+Qwen3-R	0.581	0.777	0.103	0.420	0.494	0.524	0.730	0.518
BM25+Rank1	0.520	0.758	0.161	0.584	0.533	0.610	0.624	0.541
Qwen3-8B	0.819	0.854	0.094	0.700	0.610	0.701	0.774	0.650

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

System	NeuCLIR	RAG	Fair Rank	CAsT	CRUX-MultiNews	CRUX-DUC04	RAGTIME	Macro avg
Qwen3-8B+Qwen3-R	0.860	0.898	0.052	0.675	0.620	0.713	0.609	0.632
Qwen3-8B+Rank1	0.821	0.915	0.124	0.704	0.616	0.754	0.800	0.676

Diversity-aware evaluation narrows several of the gaps observed under plain relevance. Table 4 shows that Qwen3-8B+Rank1 remains the strongest macro system at alpha-nDCG@20 = 0.543, but Qwen3-8B (0.522), Qwen3-8B+Qwen3-R (0.521), and BM25+Rank1 (0.497) are closer than under nDCG. The BM25+Rank1 configuration is particularly informative: its lexical first stage is weaker on macro nDCG, yet reranking allows it to surface complementary evidence across several datasets.

Table 4. alpha-nDCG@20 for all datasets and retrieval configurations.

System	NeuCLIR	RAG	Fair Rank	CAsT	CRUX-MultiNews	CRUX-DUC04	RAGTIME	Macro avg
BM25	0.349	0.450	0.109	0.357	0.469	0.476	0.486	0.385
BM25+Qwen3-R	0.583	0.633	0.065	0.321	0.551	0.554	0.518	0.461
BM25+Rank1	0.538	0.658	0.122	0.429	0.583	0.621	0.527	0.497
Qwen3-8B	0.627	0.683	0.074	0.437	0.648	0.652	0.531	0.522
Qwen3-8B+Qwen3-R	0.691	0.705	0.036	0.407	0.652	0.663	0.490	0.521
Qwen3-8B+Rank1	0.613	0.742	0.090	0.440	0.634	0.690	0.590	0.543

Table 5 reports the target metric. Qwen3-8B+Rank1 has the highest macro StRecall@20 at 0.702, followed by Qwen3-8B at 0.683 and Qwen3-8B+Qwen3-R at 0.672. BM25+Rank1 reaches 0.639, BM25+Qwen3-R reaches 0.617, and BM25 reaches 0.565. The macro ordering broadly follows relevance, but the per-dataset values reveal cases in which the rankings diverge.

Table 5. StRecall@20 for all datasets and retrieval configurations.

System	NeuCLIR	RAG	Fair Rank	CAsT	CRUX-MultiNews	CRUX-DUC04	RAGTIME	Macro avg
BM25	0.545	0.708	0.171	0.577	0.634	0.659	0.664	0.565
BM25+Qwen3-R	0.684	0.819	0.143	0.576	0.689	0.696	0.715	0.617
BM25+Rank1	0.664	0.818	0.217	0.639	0.710	0.732	0.690	0.639
Qwen3-8B	0.836	0.899	0.117	0.597	0.813	0.814	0.707	0.683

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

System	NeuCLIR	RAG	Fair Rank	CAsT	CRUX-MultiNews	CRUX-DUC04	RAGTIME	Macro avg
Qwen3-8B+Qwen3-R	0.839	0.903	0.060	0.600	0.817	0.826	0.660	0.672
Qwen3-8B+Rank1	0.786	0.935	0.133	0.619	0.831	0.839	0.773	0.702

Figures 3 and 4 show the relationship between the two predictor metrics and coverage. In Figure 3, nDCG@20 is positively associated with StRecall@20, but observations with similar relevance can differ meaningfully in covered nuggets. Figure 4 is tighter because alpha-nDCG discounts redundant subtopic gain. The diagonal is a scale reference rather than an equality expectation: StRecall counts whether a nugget appears at least once, whereas nDCG and alpha-nDCG include rank discounting and graded gain.

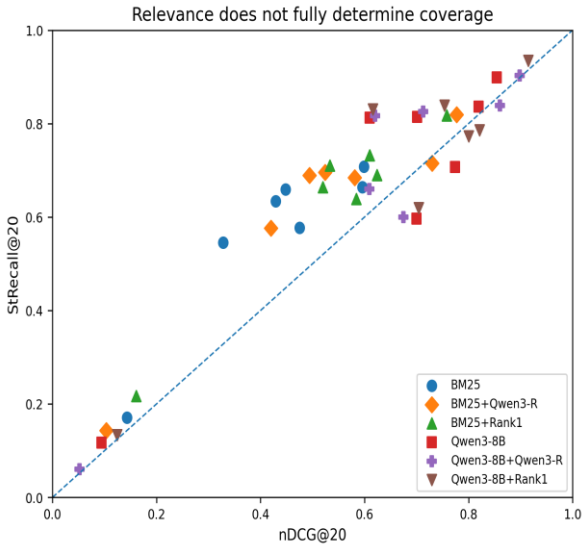

Figure 3. nDCG@20 versus StRecall@20 across all system-dataset observations.

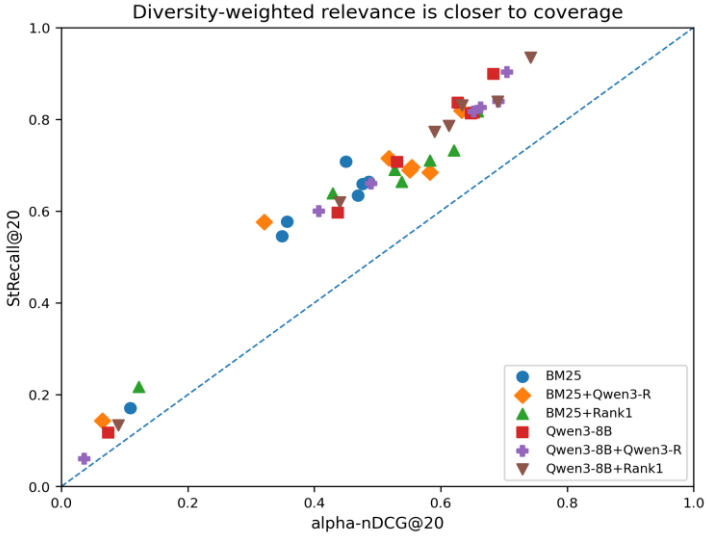

Figure 4. alpha-nDCG@20 versus StRecall@20 across all system-dataset observations.

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

4.3 Coverage-prediction accuracy

The leave-one-dataset-out comparison in Table 6 shows that NLCP-Ridge provides the most accurate held-out estimates. It obtains MAE = 0.0464, RMSE = 0.0687, R2 = 0.9090, Pearson r = 0.9628, and Spearman rho = 0.9713. The nDCG-only model has MAE = 0.1121; NLCP-Ridge therefore reduces absolute error by 58.6%. The alpha-nDCG-only model reaches MAE = 0.0566, and the proposed model improves on it by 18.0%. The nDCG+alpha baseline is strong at MAE = 0.0474, while reranker identity lowers error by a further 2.1%.

Table 6. Leave-one-dataset-out StRecall@20 prediction accuracy.

Predictor	MAE	RMSE	R2	Pearson r	Spearman rho
NLCP-Ridge	0.046	0.069	0.909	0.963	0.971
nDCG+alpha	0.047	0.074	0.894	0.965	0.968
alpha-nDCG-only	0.057	0.080	0.878	0.949	0.968
nDCG-only	0.112	0.145	0.592	0.786	0.789
System priors	0.187	0.260	-0.308	-0.495	-0.124
Mean-only	0.188	0.263	-0.330	-0.956	-0.798
Dataset priors	0.357	0.429	-2.554	-0.570	-0.569

Figure 5 makes the error reduction visible. Metric-based models dominate prior-only baselines, and alpha-nDCG is substantially more informative than nDCG alone. On a hypothetical topic inventory of 20 equally weighted nuggets, an MAE of 0.046 corresponds to less than one nugget on average, whereas an MAE of 0.112 corresponds to more than two. The comparison is approximate because the reported target is an aggregate score, but it illustrates the practical scale of the improvement.

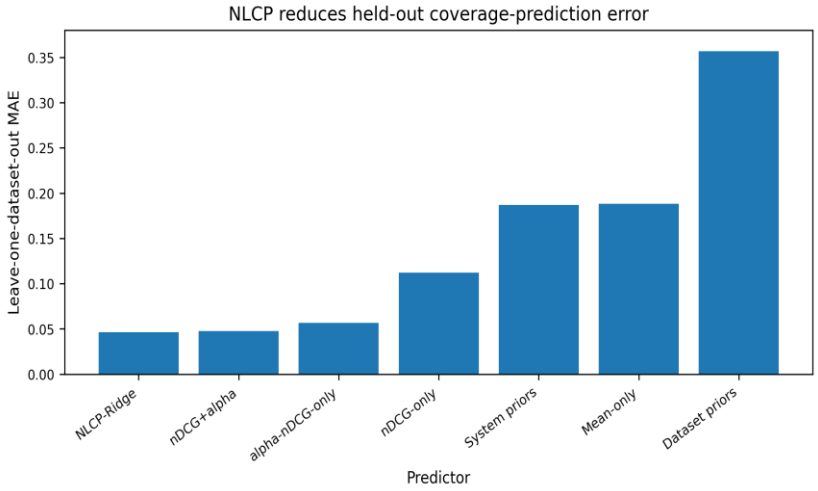

Figure 5. Mean absolute prediction error by predictor.

4.4 Dataset transfer and signal analysis

Table 7 decomposes NLCP-Ridge error by held-out dataset. RAGTIME has the lowest MAE at 0.0142, followed by RAG at 0.0195, CRUX-MultiNews at 0.0201, and CRUX-DUC04 at 0.0247. NeuCLIR remains accurate at 0.0382. Error increases for CAsT to 0.0613 and is highest for Fair Ranking at 0.1471. Figure 6 shows the same progression and identifies Fair Ranking as the primary distribution-shift case.

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

Table 7. Dataset-wise NLCP-Ridge errors under leave-one-dataset-out validation.

Dataset	MAE	RMSE	Best error	Worst error	Queries	Avg nuggets	Docs
CAsT	0.061	0.073	0.021	0.141	25	6.1	38,429,852
CRUX-DUC04	0.025	0.028	0.002	0.042	50	7.8	565,015
CRUX-MultiNews	0.020	0.028	0.001	0.061	100	14.2	565,015
Fair Rank	0.147	0.150	0.099	0.179	50	29.7	6,475,537
NeuCLIR	0.038	0.047	0.006	0.070	19	14.9	10,038,768
RAG	0.019	0.034	0.001	0.082	56	13.9	113,520,750
RAGTIME	0.014	0.019	0.003	0.042	34	15.6	4,000,380

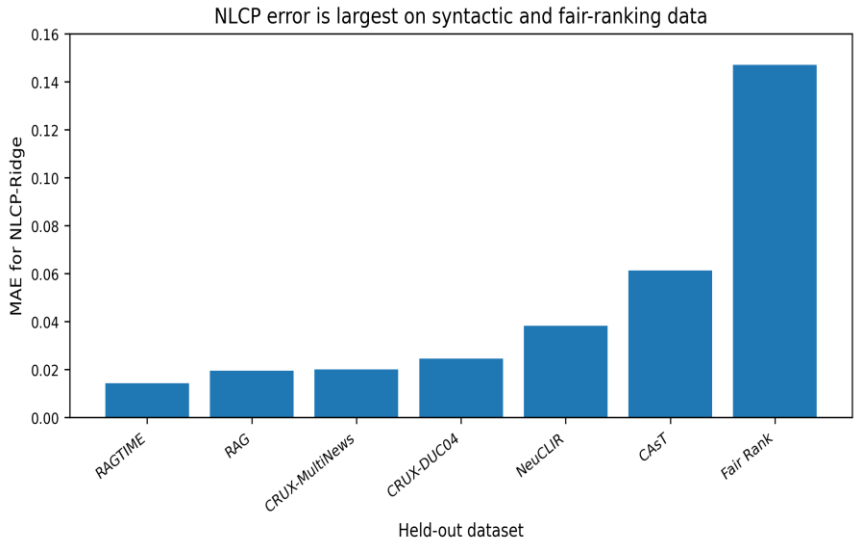

Figure 6. Dataset-wise NLCP-Ridge prediction error.

The correlation analysis in Table 8 explains why alpha-nDCG carries most of the predictive signal. Across all 42 observations, alpha-nDCG@20 has Pearson r = 0.981 and Spearman rho = 0.968 with StRecall@20. nDCG@20 remains strongly related to coverage, with Pearson r = 0.926 and Spearman rho = 0.848, but the association is weaker. The relevance-redundancy gap and alpha-to-nDCG ratio have only modest correlations, indicating that the absolute diversity-aware score is more useful than either derived contrast on this benchmark.

Table 8. Correlations between metric signals and StRecall@20.

Signal	Pearson r	Pearson p	Spearman rho	Spearman p	Kendall tau
nDCG@20	0.926	<0.001	0.848	<0.001	0.700
alpha-nDCG@20	0.981	<0.001	0.968	<0.001	0.872
Relevance-redundancy gap	0.238	0.129	0.198	0.210	0.135
Alpha-to-nDCG ratio	0.335	0.030	0.315	0.042	0.171

Macro averages in Table 9 show that Qwen3-8B+Rank1 ranks first under both nDCG@20 and StRecall@20, Qwen3-8B ranks second, and Qwen3-8B+Qwen3-R ranks third. This agreement at the macro level is useful but

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

incomplete: averaging conceals the dataset-level disagreements examined below. The practical implication is that a globally strong retriever can still be suboptimal for coverage on a particular information need.

Table 9. Macro-average retrieval and coverage performance by system.

System	nDCG@20	alpha-nDCG@20	StRecall@20	StRecall rank	nDCG rank
BM25	0.431	0.385	0.565	6	6
BM25+Qwen3-R	0.518	0.461	0.617	5	5
BM25+Rank1	0.541	0.497	0.639	4	4
Qwen3-8B	0.650	0.522	0.683	2	2
Qwen3-8B+Qwen3-R	0.632	0.521	0.672	3	3
Qwen3-8B+Rank1	0.676	0.543	0.702	1	1

Table 10 confirms that the compact feature design is important. Dataset priors alone perform poorly because collection descriptors do not reveal how a system ranked the evidence. System priors also fail to capture query and collection difficulty. nDCG-only improves substantially, alpha-nDCG-only improves further, and the fused metric baseline nearly matches the proposed model. The final gain from reranker identity is small but consistent, suggesting that reranking family retains information about coverage behavior after the two metric scores are observed.

Table 10. Predictor ablation results relative to NLCP-Ridge.

Predictor	MAE	RMSE	R2	Pearson r	Spearman rho	Delta MAE
NLCP-Ridge	0.046	0.069	0.909	0.963	0.971	0.000
nDCG+alpha	0.047	0.074	0.894	0.965	0.968	0.001
alpha-nDCG-only	0.057	0.080	0.878	0.949	0.968	0.010
nDCG-only	0.112	0.145	0.592	0.786	0.789	0.066
System priors	0.187	0.260	-0.308	-0.495	-0.124	0.141
Mean-only	0.188	0.263	-0.330	-0.956	-0.798	0.142
Dataset priors	0.357	0.429	-2.554	-0.570	-0.569	0.310

4.5 Winner disagreements and operational implications

Table 11 provides the clearest evidence that relevance-only model selection can miss the best coverage policy. Five datasets have the same nDCG and StRecall winner, but two do not. On CAsT, Qwen3-8B+Rank1 has the highest nDCG@20 at 0.704, whereas BM25+Rank1 has the highest StRecall@20 at 0.639. On CRUX-MultiNews, Qwen3-8B+Qwen3-R leads nDCG@20 at 0.620, while Qwen3-8B+Rank1 leads StRecall@20 at 0.831. Alpha-nDCG selects the same relevance winner in both disagreement cases, showing that even a diversity-aware ranked metric does not guarantee the largest final nugget union.

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

Table 11. Winner disagreement across relevance, diversity, and coverage metrics.

Dataset	Best nDCG system	Best nDCG	Best alpha system	Best alpha	Best StRecall system	Best StRecall	nDCG=St Recall	alpha=St Recall
CAsT	Qwen3-8B+Rank1	0.704	Qwen3-8B+Rank1	0.440	BM25+Rank1	0.639	No	No
CRUX-DUC04	Qwen3-8B+Rank1	0.754	Qwen3-8B+Rank1	0.690	Qwen3-8B+Rank1	0.839	Yes	Yes
CRUX-MultiNews	Qwen3-8B+Qwen3-R	0.620	Qwen3-8B+Qwen3-R	0.652	Qwen3-8B+Rank1	0.831	No	No
Fair Rank	BM25+Rank1	0.161	BM25+Rank1	0.122	BM25+Rank1	0.217	Yes	Yes
NeuCLIR	Qwen3-8B+Qwen3-R	0.860	Qwen3-8B+Qwen3-R	0.691	Qwen3-8B+Qwen3-R	0.839	Yes	Yes
RAG	Qwen3-8B+Rank1	0.915	Qwen3-8B+Rank1	0.742	Qwen3-8B+Rank1	0.935	Yes	Yes
RAGTIME	Qwen3-8B+Rank1	0.800	Qwen3-8B+Rank1	0.590	Qwen3-8B+Rank1	0.773	Yes	Yes

Coverage gaps in Table 12 further show that the relationship is not a fixed offset. BM25 has an average StRecall-minus-nDCG gap of 0.134, while Qwen3-8B+Rank1 has a much smaller average gap of 0.026. Even the best macro system ranges from a maximum positive gap of 0.215 to a minimum gap of -0.085 across datasets. Positive StRecall-minus-alpha gaps occur for every system because StRecall counts binary nugget presence, whereas alpha-nDCG discounts position and repeated gain.

Table 12. Coverage-gap summary by system.

System	Mean StRecall-nDCG	Mean StRecall-alpha	Max StRecall-nDCG	Min StRecall-nDCG
BM25	0.134	0.180	0.217	0.028
BM25+Qwen3-R	0.099	0.157	0.195	-0.015
BM25+Rank1	0.097	0.142	0.177	0.055
Qwen3-8B	0.033	0.162	0.203	-0.103
Qwen3-8B+Qwen3-R	0.040	0.152	0.197	-0.075
Qwen3-8B+Rank1	0.026	0.160	0.215	-0.085

Figure 7 displays the full StRecall@20 matrix and highlights three distinct regimes: high coverage on RAG and CRUX for neural retrieval, low coverage across all methods on Fair Ranking, and a comparatively narrow ceiling on CAsT. A deployment policy should therefore treat predicted coverage as query- and collection-sensitive rather than assuming that the same retriever is uniformly best.

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

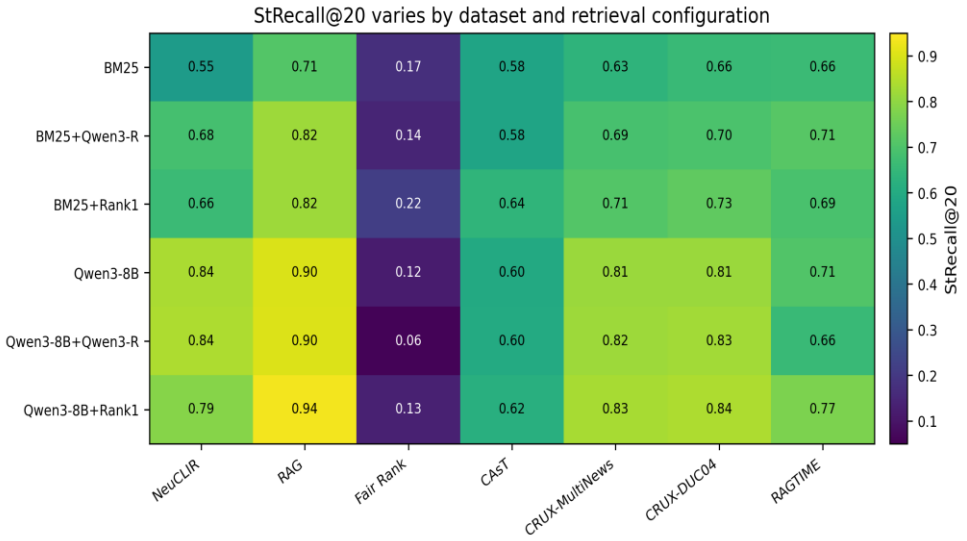

Figure 7. StRecall@20 heatmap over systems and datasets.

Figure 8 compares observed and predicted StRecall@20. Most observations lie close to the diagonal, with the largest deviations concentrated in Fair Ranking. The predicted score can be converted into an intuitive completeness-risk signal, 1 - predicted StRecall@20. A high risk can trigger additional lexical retrieval, subquery generation, MMR-style diversification, a larger candidate budget, or selective abstention. The predictor does not prescribe one remedy; it identifies when the current evidence set is unlikely to be broad enough.

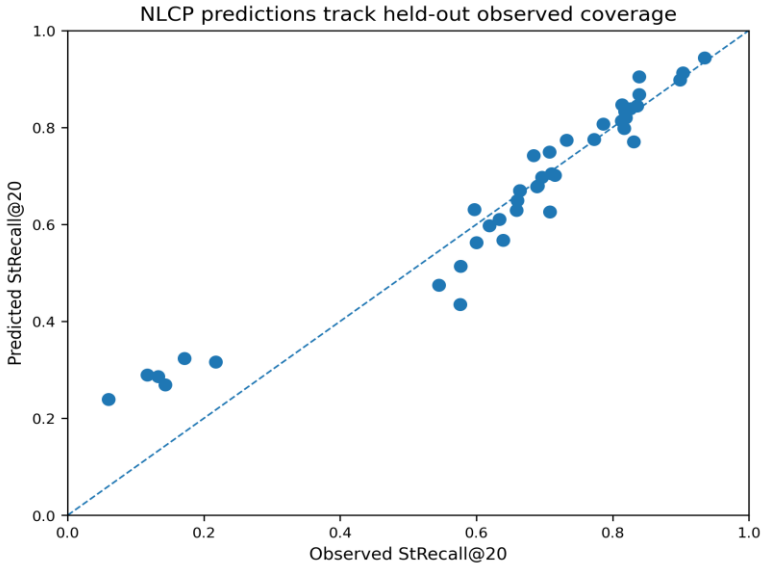

Figure 8. Observed versus predicted StRecall@20 for NLCP-Ridge.

The results support four conclusions. First, relevance and coverage are strongly related but not interchangeable. Second, diversity-aware relevance is a better coverage proxy than plain nDCG. Third, the addition of reranker identity improves transfer slightly after both metric signals are known. Fourth, dataset-level analysis remains necessary because Fair Ranking, CAsT, and CRUX-MultiNews expose different forms of coverage mismatch. In an end-to-end RAG system, these findings motivate a two-stage control strategy: assess whether retrieval is broad enough before generation, then verify whether generated claims remain supported afterward.

5. Limitations

The prediction target is an aggregate StRecall@20 value for each system-dataset pair. This level of analysis is appropriate for comparing retrieval configurations and testing cross-collection transfer, but it does not produce a

ISSN: 3071-4656
DOI: 10.66372/JGER.V4I2.4

calibrated coverage probability for an individual query or nugget. Query-level modeling would require topic-level rankings and nugget judgments, along with features that describe facet count, first-hit positions, passage overlap, and redundancy within each retrieved set.

The dataset contains 42 observations, so the model is deliberately linear and compact. More flexible nonlinear estimators could fit the observed matrix more closely but would be difficult to validate reliably under full-dataset holdout. The strong error on Fair Ranking also shows that a collection with unusual lexical and nugget structure can fall outside the relationship learned from the other datasets.

The study uses a fixed cutoff of 20. Coverage at rank 5, rank 50, or under a token budget may lead to different operating points because documents vary in length and nugget density. The same framework can be extended to other cutoffs, but the fitted coefficients and intervention thresholds should not be assumed to transfer unchanged.

Finally, predicted coverage is not a complete faithfulness evaluation. It estimates whether the retrieved set contains the target information units; it does not prove that a generator will cite the correct passage, preserve numerical values, resolve conflicting evidence, or avoid unsupported synthesis. Coverage prediction should therefore be paired with attribution and claim-level verification rather than treated as a substitute for them.

6. Conclusion

This study examined nugget-level retrieval coverage prediction across seven CoverageBench datasets and six retrieval configurations. NLCP-Ridge combines nDCG@20, alpha-nDCG@20, and reranker identity in a deterministic ridge model evaluated under leave-one-dataset-out validation. Across 42 held-out cases, the model achieves MAE = 0.0464 and R2 = 0.9090, substantially outperforming relevance-only and prior-only baselines.

The retrieval comparisons show why a separate coverage diagnostic is needed. Strong relevance usually accompanies strong coverage, but CAsT and CRUX-MultiNews select different system winners under nDCG and StRecall, and Fair Ranking exhibits a distinct transfer pattern. Alpha-nDCG is the strongest single proxy because it accounts for repeated subtopic gain, yet the final nugget union still contains information that ranked relevance metrics do not fully capture.

For RAG design, the central implication is operational: evidence should be assessed not only for relevance and support, but also for breadth. Predicted StRecall can guide additional retrieval, diversification, or abstention before generation. Future work should extend the analysis to query-level coverage with uncertainty intervals and then test whether retrieval-side risk estimates improve answer-level nugget recall, citation quality, and user judgments of completeness.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. NeurIPS, 2020.

[2] 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.

[3] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in Proc. EACL, 2021, pp. 874–880.

[4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” arXiv:2312.10997, 2023.

[5] 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.

[6] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated evaluation of retrieval augmented generation,” in Proc. EACL System Demonstrations, 2024, pp. 150–158.

[7] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, “ARES: An automated evaluation framework for retrieval-augmented generation systems,” in Proc. NAACL-HLT, 2024, pp. 338–354.

[8] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, “FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in Proc. EMNLP, 2023, pp. 12076–12100.

[9] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in Proc. SIGIR, 1998, pp. 335–336.

[10] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, “Novelty and diversity in information retrieval evaluation,” in Proc. SIGIR, 2008, pp. 659–666.

[11] C. L. A. Clarke, N. Craswell, and I. Soboroff, “Overview of the TREC 2009 Web Track,” in Proc. TREC, 2009.

[12] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S. Wu, “Intent-based diversification of web search results: Metrics and algorithms,” Inf. Retr., vol. 14, no. 6, pp. 572–592, 2011.

[13] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, “A comparative analysis of cascade measures for novelty and diversity,” in Proc. WSDM, 2011, pp. 75–84.

[14] E. M. Voorhees, “Overview of the TREC 2003 question answering track,” in Proc. TREC, 2003.

[15] A. Nenkova and R. Passonneau, “Evaluating content selection in summarization: The pyramid method,” in Proc. HLT-NAACL, 2004, pp. 145–152.

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

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

[18] A. Moffat and J. Zobel, “Rank-biased precision for measurement of retrieval effectiveness,” ACM Trans. Inf. Syst., vol. 27, no. 1, pp. 1–27, 2008.

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

[20] T. Formal, B. Piwowarski, and S. Clinchant, “SPLADE: Sparse lexical and expansion model for first stage ranking,” in Proc. SIGIR, 2021, pp. 2288–2292.

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

[22] R. Nogueira and K. Cho, “Passage re-ranking with BERT,” arXiv:1901.04085, 2019.

[23] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, “CEDR: Contextualized embeddings for document ranking,” in Proc. SIGIR, 2019, pp. 1101–1104.

[24] J. Lin, X. Ma, S. Lin, J. Yang, R. Pradeep, and R. Nogueira, “Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations,” in Proc. SIGIR, 2021, pp. 2356–2362.

[25] 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

[26] S. Samuel, A. Yates, D. Lawrie, I. Soboroff, T. Adriaanse, B. Van Durme, and E. Yang, “CoverageBench: Evaluating information coverage across tasks and domains,” arXiv:2603.20034, 2026.

ISSN: 3071-4656

DOI: 10.66372/JGER.V4I2.4

[27] R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, and J. Lin, "The Great Nugget Recall: Automating fact extraction and RAG evaluation with large language models," in Proc. SIGIR, 2025, pp. 180–190, doi: 10.1145/3726302.3730090.

[28] J.-H. Ju, F. G. Landry, E. Yang, S. Verberne, and A. Yates, "LANCER: LLM reranking for nugget coverage," arXiv:2601.22008, 2026.

[29] J.-H. Ju, E. Yang, T. Adriaanse, S. Verberne, and A. Yates, "Search for Coverage: Learning coverage-aware retrieval with augmented sub-question answerability," arXiv:2605.28522, 2026.

[30] 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.

[31] 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.

[32] 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.

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

[34] G. Liu, C. Li, and E. Zhang, "OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations," JACS, vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

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

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

[37] 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.

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

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

[40] 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.

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

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

[43] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” J. R. Stat. Soc. B, vol. 36, no. 2, pp. 111–147, 1974.

[44] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970.