RARE: Redundancy-Aware Retrieval Evaluation Framework for High-Similarity Corpora

Anonymous ACL submission

Abstract

Existing QA benchmarks typically assume distinct documents with minimal overlap, yet real-world retrieval-augmented generation (RAG) systems operate on corpora such as financial reports, legal codes, and patents, where information is highly redundant and documents exhibit strong cross-document similarity. This mismatch undermines evaluation validity: retrievers may surface correct evidence yet be penalized because redundancy across documents is not accounted for, or may score well despite failing to transfer to deployment settings. We present **RARE** (Redundancy-Aware Retrieval Evaluation), a framework for constructing realistic benchmarks by (i) decomposing documents into atomic facts to enable precise redundancy tracking, and (ii) stabilizing LLM-based question generation through **SepRRF**, which evaluates quality criteria independently and fuses rankings rather than relying on unstable confidence scores. Applying RARE to finance, law, and patent corpora, we introduce **RedQA**, where state-of-the-art retrievers drop from 78% on HotpotQA to 5.0–31.7% Top@10 accuracy at 4-hop depth, revealing robustness gaps that current benchmarks fail to capture. RARE enables practitioners to build domain-specific RAG evaluations that faithfully reflect real-world deployment conditions.

1 Introduction

Large language models (LLMs) require external knowledge retrieval to remain accurate and current (Lewis et al., 2020; Izacard et al., 2022). In enterprise settings, retrieval-augmented generation (RAG) operates on specialized corpora—financial disclosures, legal statutes, patent filings—that exhibit a critical property: **high information redundancy and inter-document similarity** (Zheng et al., 2024).

Standard benchmarks, however, do not reflect these conditions. As Figure 1 shows, HotpotQA (Yang et al., 2018) exhibits 1.4% redundancy and 8.8% similarity, whereas finance (63.2%, 35.1%), legal (25.1%, 40.7%), and patent (49.7%, 29.0%) corpora show substantially higher overlap. Here, **redundancy** measures the proportion of atomic facts that appear in multiple documents, and **similarity** measures the average cosine similarity between *passage* embeddings (Thakur et al., 2021). This mismatch yields two evaluation failures. First, in high-redundancy settings, retrievers that correctly *surface alternative evidence* are penalized by *single-canonical-passage* labeling (Wu et al., 2024). Second, because widely used benchmarks are *low-similarity* while real deployments are *high-similarity*, This overestimates deployment performance; under high similarity, dense retrievers degrade more than sparse ones (§ 4.2).

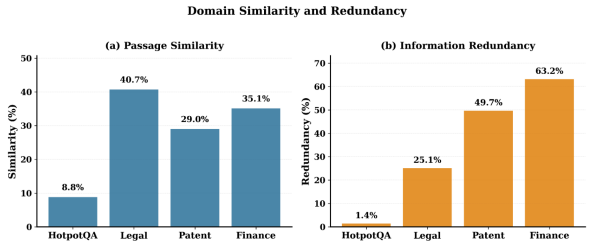

Figure 1: Passage-level similarity and information redundancy across domains. HotpotQA shows minimal overlap, whereas Finance, Patent, and Legal exhibit substantially higher levels.

Constructing benchmarks that reflect these conditions requires two capabilities. First, **high-quality question generation**: LLMs frequently produce trivial, unnatural, or contradictory questions (Liu et al., 2023; Pan et al., 2023) (Figure 3; Appendix F), degrading dataset validity. Second, **systematic redundancy tracking**: evaluation must identify where the same information recurs across the dataset (Voorhees, 2000; Zheng et al., 2024). Unlike training data, which demands only local control over positive/negative pairs (Karpukhin

et al., 2020), evaluation data must account for identical evidence appearing in multiple passages; otherwise, retrievers surfacing valid alternatives are unfairly penalized (Wu et al., 2024).

We introduce **RARE** (Redundancy-Aware Retrieval Evaluation), a modular framework that addresses both challenges. RARE decomposes documents into atomic facts, tracks redundancy via embedding similarity combined with LLM verification, and stabilizes question generation with **SepRRF** (Separate Prompting with Reciprocal Rank Fusion), which evaluates quality criteria independently and aggregates results via rank-based fusion that is less sensitive to poorly calibrated confidence scores (Cormack and Clarke, 2009; Wang et al., 2023).

Experiments reveal severe degradation under realistic conditions. For 4-hop queries across the Finance, Legal, and Patent domains, state-of-the-art retrievers attain at most **31.7% Top@10**, with Finance and Legal dropping to **8.5%** and **5.0%**, respectively, whereas HotpotQA maintains **77.9%** at 4-hop (Yang et al., 2018).

In our evaluations (§ 4.2), dense retrievers exhibit substantial degradation in high-similarity corpora—e.g., the Coverage@10 gap between BM25 and a strong dense retriever shrinks from 28.2%p on HotpotQA to only 7.7%p in Finance and 12.1%p in Legal. This pattern reveals that semantic retrieval, while highly effective on low-overlap benchmarks, can lose much of its advantage in realistic, similarity-rich domains, underscoring the need for evaluation frameworks that accurately capture these retrieval conditions.

Contributions. Our contributions are threefold:

- We quantify the gap between academic benchmarks and enterprise corpora via redundancy and similarity analysis, and introduce **RedQA**, a benchmark capturing these conditions across finance, law, and patents.
- We trace question-generation failures to complex multi-criterion reasoning and introduce **SepRRF**, which achieves higher quality and stability through independent criterion evaluation and rank fusion.
- We release **RARE**, a generalizable framework enabling practitioners to construct domain-specific RAG benchmarks that reflect deployment realities.

2 RARE Framework

2.1 Pipeline Overview

RARE constructs realistic multi-hop RAG evaluation data under high redundancy and similarity. Figure 2 illustrates the pipeline, which consists of three stages: *Valid Information Selection*, *Redundancy Tracking*, and *Question-Answer Generation*. Each selection stage follows a three-step pattern: construct a candidate pool, filter invalid candidates, and rank by multi-criteria preferences.

A central component is our ranking method, **SepRRF** (detailed in §2.5), which stabilizes LLM judgments by evaluating each quality criterion independently and fusing results through rank aggregation rather than confidence scores. By decomposing complex multi-criterion reasoning into simple binary or scalar judgments, SepRRF produces more consistent and human-aligned decisions.

Notation.
- C: input document chunk
- $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$: set of extracted atomic information units
- $\mathcal{A}_{\text{valid}} \subseteq \mathcal{A}$: atomic units passing validity filters
- $\mathcal{T} \subseteq \mathcal{A}_{\text{valid}}$: top-$k$ target units selected per chunk
- $\phi(\cdot) \in \{\text{PASS}, \text{FAIL}\}$: binary LLM judge
- $\sigma(\cdot) \in [0, 1]$: scalar LLM scoring function

2.2 Valid Information Selection

This stage builds a reliable pool of candidate facts for multi-hop question generation through three steps: (i) decompose chunks into minimal factual claims, (ii) filter units that fail minimal validity, and (iii) prioritize the remainder by multi-criteria ranking.

Atomic Information Extraction. Each input chunk C is decomposed by an LLM into atomic information units:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_n\} = f_{\text{LLM}}(C).$$

Each a_i denotes the smallest indivisible factual statement from the text. Unlike passages where multiple facts intertwine, atomic units isolate single claims. This granularity enables (i) precise cross-document redundancy tracking and (ii) modular building blocks for flexible multi-hop composition. Refer to Appendix B and Appendix F.1 for the actual examples and prompt design.

Validity Filtering. Not all extracted units are suitable for retrieval-based QA. We remove units that fail minimal validity:

$$\mathcal{A}_{\text{valid}} = \{\, a \in \mathcal{A} \mid \phi(a) = \text{PASS} \,\}.$$

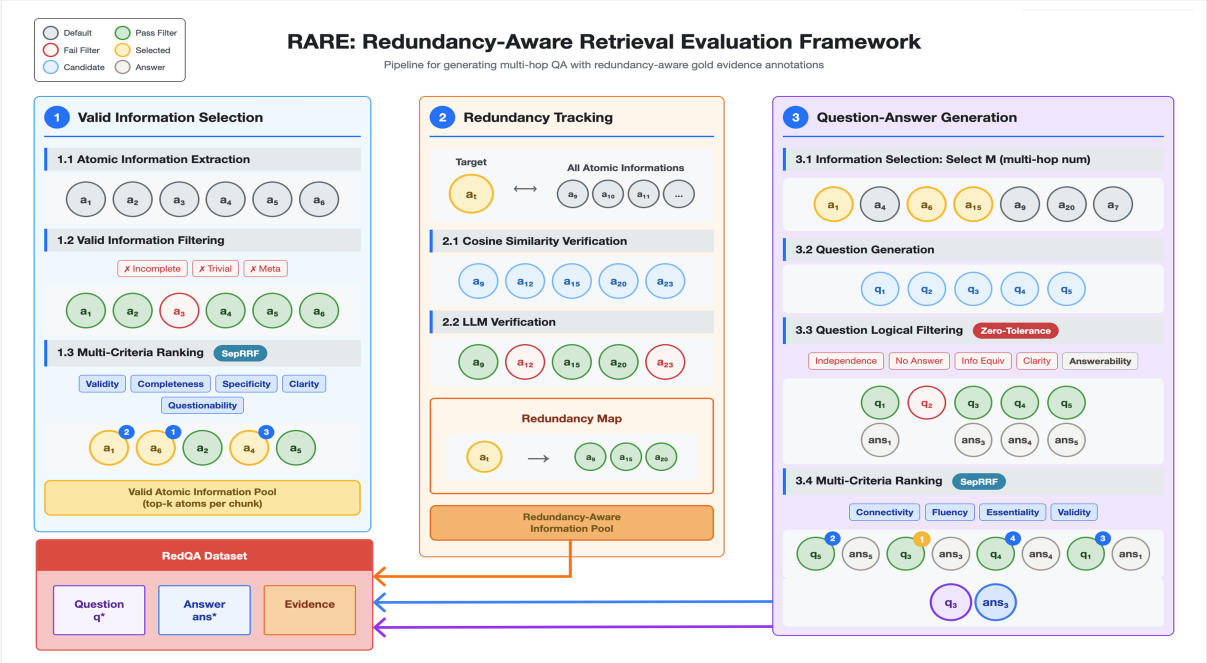

Figure 2: Overall workflow of the **RARE** framework. The pipeline consists of valid information selection, redundancy tracking, and question-answer generation. Multi-criteria judgments are stabilized through SepRRF, and redundancy-aware labeling ensures fair evaluation under real-world corpus conditions. During the logical filtering step in Question-Answer Generation, answerability validation additionally performs answer generation.

Each unit is checked against three minimal criteria: (1) the unit is complete and self-contained, (2) it conveys non-trivial, utility-relevant knowledge, and (3) it expresses factual content rather than metadata or structural fragments. This step prunes incomplete or low-value units, leaving a pool of reliable factual knowledge. (Appendix F.2)

Multi-Criteria Ranking. Passing minimal validity does not make all remaining units equally useful. We assess each atomic unit on five criteria (Appendix F.3): (1) *Validity*—whether the fact is substantively meaningful, (2) *Completeness*—whether it is interpretable without external context, (3) *Specificity*—whether it provides concrete details (names, dates, numbers), (4) *Clarity*—whether it avoids ambiguity, and (5) *Questionability*—whether it can naturally form a question. Criterion-wise ranks are aggregated via SepRRF (§2.5). From each chunk we select the top-k units to construct the *Valid Information Pool*, supporting redundancy tracking and subsequent question generation.

2.3 Systematic Redundancy Tracking

Problem Definition. In real corpora, identical factual content often appears across multiple chunks. Suppose question q_i is generated from source chunk P_i. While the answer is contained in the source, $\mathrm{Ans}(q_i) \subseteq P_i$, it may also be supported by other chunks: $\exists j \neq i$ such that $\mathrm{Ans}(q_i) \subseteq P_j$. A retriever returning such P_j is factually correct but would be penalized under naive gold annotation that assumes a single origin for each answer.

Two-Stage Detection. Starting from the *Valid Information Pool*, we select the top-k atomic units from each chunk to form the target set $\mathcal{T} \subseteq \mathcal{A}_{\text{valid}}$. For every target $a_t \in \mathcal{T}$, we first retrieve a high-recall candidate set using embedding similarity:

$$\mathcal{C}_\tau(a_t) = \{\, a_j \in \mathcal{A} \mid \mathrm{sim}(a_t, a_j) \geq \tau \}$$

where $\mathrm{chunk}(a_t)$ denotes the originating chunk. We then refine candidates with an LLM Judge $\phi(a_t, a_j) \in \{\textsc{Equiv}, \textsc{NonEquiv}\}$:

$$\mathcal{R}(a_t) = \{\, a_j \in \mathcal{C}_\tau(a_t) \mid \phi(a_t, a_j) = \textsc{Equiv} \,\}.$$

Atomic granularity reduces noise in the embedding space and narrows the gap between semantic similarity and factual equivalence, making equivalence checks more tractable for LLMs. Embedding similarity search uses a recall-focused threshold ($\tau = 0.5$) to emphasize recall, while LLM verification enforces precision. Redundancy labels are recorded as key–value mappings $(a_t \mapsto \mathcal{R}(a_t))$ for all targets.

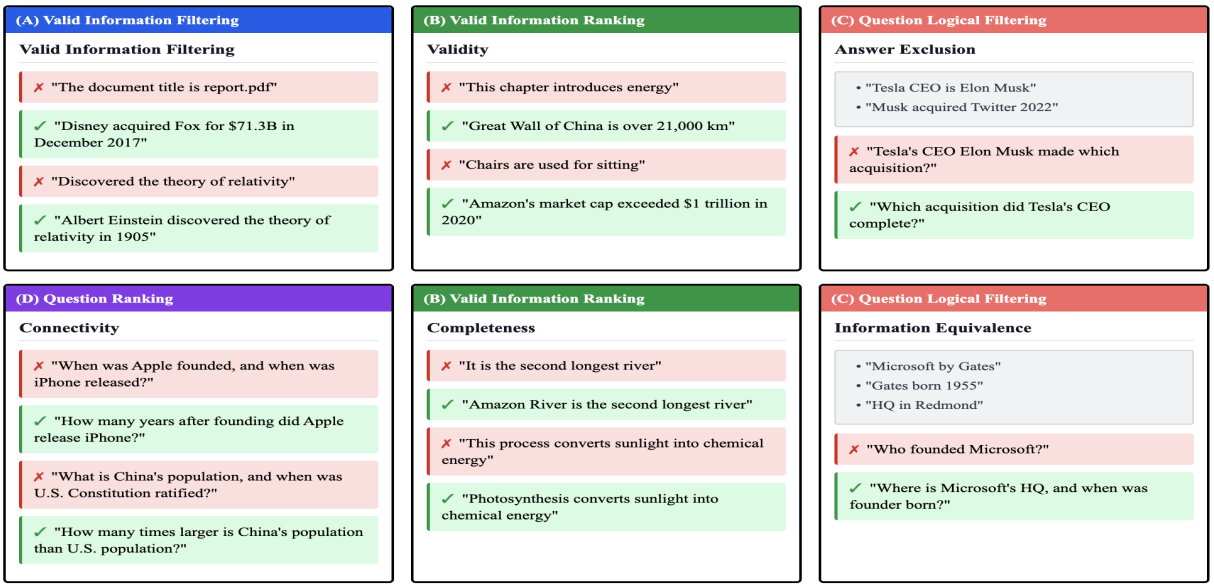

Figure 3: Representative examples of filtering and ranking criteria in RARE. (A) Valid information filtering removes incomplete units lacking essential elements. (B) Two of five quality criteria for information ranking: Validity (substantive usefulness) and Completeness (self-contained clarity). (C) Two of five logical checks for questions: Answer Exclusion (no embedded answers) and Information Equivalence (exact scope matching). (D) One of four quality criteria for question ranking: Connectivity (natural logical flow). Grey boxes indicate given information required for evaluation. Complete criterion descriptions and additional examples are provided in Appendix F.

2.4 Question-Answer Generation

Candidate facts are combined into reasoning chains, filtered for logical soundness, and ranked by multi-criterion judgments to yield challenging yet well-formed multi-hop items.

Question Candidate Construction. From the redundancy-aware valid pool, we randomly sample N atomic units to form input set $A_N \subset \mathcal{A}_{\text{valid}}$. Within this set, the LLM selects exactly M units with strong relational connectivity, encouraging connected reasoning rather than mere fact listing:

$$T = f_{\text{select}}(A_N, M), \quad T \subseteq A_N, \ |T| = M.$$

Using these M units as common evidence, we generate S paraphrased candidate questions that all require the same underlying evidence:

$$Q = \{\, q_1, q_2, \ldots, q_S \,\} = f_{\text{gen}}(T, S).$$

Producing multiple candidates mitigates generation-time errors by enabling later filtering and ranking to discard flawed questions while retaining high-quality ones. Here, N denotes the number of sampled atoms, M the multi-hop number, and S the number of candidate questions. (Appendix A.2)

Logical Filtering (Zero-Tolerance). Because LLMs may introduce diverse logical errors during generation, we apply strict filtering based on five criteria (Appendix F.6): (1) *Contextual Independence*—the question must not rely on document structure or meta-references, (2) *Answer Exclusion*—no answer content or intermediate reasoning may be embedded in the question, (3) *Information Equivalence*—the informational scope of the question must *exactly* match the designated atomic evidence set (requiring all and only the given units), (4) *Question Clarity*—the formulation must avoid ambiguity or underspecified references, and (5) *Answerability*—the selected evidence must be substantively sufficient to yield a clear and unique answer.

This is enforced under a **zero-tolerance policy**: if a candidate fails even one criterion, it is immediately discarded. Formally, given $Q = \{q_1, \ldots, q_S\}$,

$$Q^* = \{\, q \in Q \mid \forall \ell \in \mathcal{L},\ \phi_\ell(q) = \textsc{Pass} \,\},$$

where $\mathcal{L}$ is the set of five criteria and ϕ_ℓ is the LLM Judge for criterion ℓ.

Multi-Criteria Ranking. To select the best question, each $q \in Q^*$ is evaluated on four criteria (Appendix F.7): (1) *Connectivity*—whether the selected facts form a coherent reasoning chain, (2)

Fluency—whether the question is expressed in natural language, (3) *Essentiality*—whether it focuses on core information without redundancy, and (4) *Validity*—whether it asks for substantively meaningful content. Criterion-wise ranks are aggregated via SepRRF (§2.5), and the highest-ranked question is retained as the benchmark item.

2.5 SepRRF: Stabilizing Multi-Criteria Ranking

Motivation. Ranking candidates by multiple quality criteria is challenging for LLMs. Jointly reasoning over criteria produces unstable outputs because it requires the model to balance competing objectives simultaneously. Moreover, LLM confidence scores are unreliable and poorly calibrated across criteria, making score-based aggregation prone to magnitude mismatches.

Method. SepRRF addresses both issues through two principles: (1) evaluate each criterion independently via separate prompts, and (2) fuse results using rank-based aggregation, discarding confidence magnitudes entirely.

Formally, given candidate x and N criteria, we obtain per-criterion ranks $\mathrm{rank}_i(x)$ through independent LLM calls. The fused score is:

$$s(x) = \sum_{i=1}^{N} \frac{1}{\mathrm{rank}_i(x)}$$

Higher $s(x)$ indicates broader inter-criterion consensus. Unlike confidence-weighted averaging, this approach relies only on ordinal preferences, which LLMs produce more reliably than calibrated probabilities.

Application in RARE. SepRRF is applied at two stages: (1) ranking atomic information units by 5 criteria (validity, completeness, specificity, clarity, questionability), and (2) ranking candidate questions by 4 criteria (connectivity, fluency, essentiality, validity). Both Section 4.4 and Appendix E demonstrate that SepRRF consistently achieves higher quality and stability.

2.6 Benchmark Construction: RedQA

The final outcome of RARE is **Redundancy-Aware QA (RedQA)**. Each benchmark item pairs the best-ranked question with its reference answer and a redundancy-aware gold evidence set. The gold set includes not only the originating passages but also all semantically equivalent passages identified through redundancy tracking, ensuring that retrievers are not penalized for retrieving correct but alternative evidence. RedQA consists of instances with three components: the top-ranked question, its reference answer, and the redundancy-aware gold evidence set.

Table 1: Dataset statistics. Enterprise corpora exhibit substantially higher redundancy and similarity than HotpotQA.

Corpus	Chunks	Redundancy (%)	Similarity (%)	Samples
Finance	3,281	63.2	35.1	645
Patent	3,339	49.7	29.0	480
Legal	4,090	25.1	40.7	605
HotpotQA	3,218	1.4	8.8	688

3 Dataset

We evaluate RARE on four domains. Finance, Patent, and Legal represent high-redundancy/high-similarity enterprise corpora, while HotpotQA serves as a low-overlap baseline. Table 1 summarizes corpus-level statistics.

Definitions. We report *Redundancy (%)* as the fraction of atomic facts that appear in more than one chunk after equivalence verification, and *Similarity (%)* as the average cosine similarity between passage embeddings.

Corpora. The **Finance** corpus consists of SEC 10-K filings (2023–2024) from six technology firms (Apple, Microsoft, Amazon, Meta, Tesla, NVIDIA). **Patent** comprises 2023–2024 USPTO documents across batteries, semiconductors, wireless, pharmaceuticals, and computing. **Legal** is drawn from the 2023 U.S. Code (general provisions, intellectual property, arbitration, contracts, public works). **HotpotQA** (Yang et al., 2018) provides Wikipedia passages as a low-redundancy, open-domain baseline. Additional corpus details appear in Appendix A.1.

Construction. We use GPT-5 Nano for LLM judgments and GPT-5 for question generation; text-embedding-3-large is used for similarity computations. Multi-hop questions span 1–4 hops. Full implementation details are in Appendix A.2.

3.1 Human Evaluation of LLM Filtering

We validated two filtering stages—*Information Filtering* and *Question Logical Filtering*—on 480 items, evenly sampled across hop depths (1–4), domains (Finance/Patent/Legal), and pass/fail strata.

Table 2: Human evaluation results. Our conservative approach prioritizes recall over precision to minimize false negatives that would compromise benchmark validity.

Stage	TP	FP	FN	TN	Prec.	Rec.	Acc.
Information Filtering	69	51	8	112	57.5%	**89.6%**	75.4%
Question Filtering	61	59	12	108	50.8%	**83.6%**	70.4%

Eight NLP researchers independently judged 60 items each using the same rubric as the LLM; about 17 items were adjudicated via group discussion. All assessments were conducted as part of their regular duties. Human judgments served as gold labels. Throughout, we treat *Fail* as the positive class.

Our filtering strategy achieves high recall (89.6% and 83.6%) at the cost of moderate precision (57.5% and 50.8%), resulting in overall accuracy of 75.4% and 70.4%. Notably, only **20 false negatives (4.2%)** occurred across all 480 samples. This design reflects a key principle for benchmark construction: *excluding some valid content is preferable to retaining logically flawed questions, as the latter would directly compromise evaluation validity*.

4 Experiments

4.1 Setup

Models. We compare a sparse retriever, **BM25** (Robertson and Zaragoza, 2009), against dense retrievers spanning both native embedding models and LLM-based variants. The commercial baseline is **OpenAI-Large** (text-embedding-3-large (OpenAI, 2024)). Open-source models include pure embedding architectures (E5-Large (Wang et al., 2022), BGE-M3 (Chen et al., 2024), Jina-v4 (Günther et al., 2024)) and LLM-backbone embeddings (E5-Mistral-7B (Wang et al., 2024), Qwen3-Embedding variants (Bai et al., 2025)). This suite enables controlled comparisons across sparse vs. dense paradigms, parameter scales, and architectural choices.

Evaluation Metrics. We use **Coverage@10** and **Top@10** as primary metrics, with NDCG@10 and MRR in Appendix D.

- **Coverage@10**: Proportion of required information in top-10 (e.g., 2/4 = 50%). Measures partial retrieval success under redundancy-aware evaluation.
- **Top@10**: Binary success requiring all information in top-10 (e.g., 2/4 = fail). Reflects whether retrieved context contains sufficient evidence for the user-facing system capability.

Both metrics use redundancy-aware gold labels.

4.2 Main Results

Table 3 presents retrieval performance across four domains with varying redundancy and similarity characteristics.

Sparse vs. Dense Retrieval. Dense retrievers outperform BM25 across all domains, but margins shrink in high-similarity corpora. In HotpotQA, Coverage@10 gaps are as large as 28.2%p (BM25: 64.58% vs OpenAI-Large: 92.78%), but narrow to 7.7%p (Finance) and 12.1%p (Legal), suggesting reduced effectiveness of semantic retrieval as passage similarity increases.

Scaling Effects. Model size correlates with retrieval quality. Qwen3 scales monotonically across domains: Top@10 rises from 43.80% (0.6B) to 48.84% (8B) in Finance, 35.37% to 41.49% in Legal, and 60.21% to 63.12% in Patent. Pure embeddings show similar trends: Jina-v4 (3.75B) outperforms E5-Large (0.56B) by 2.33%p in Finance Top@10 and 3.64%p in Legal Top@10. However, even the largest state-of-the-art models achieve only 41.49% (Legal) and 48.84% (Finance) at Top@10, revealing severe performance limitations in high-similarity domains.

Architectural Effects. LLM-based embeddings consistently outperform pure embedding models at comparable scales. At 0.6B parameters in Legal, Qwen3-0.6B attains 60.29% Coverage@10 versus 52.56% (E5-Large, +7.73%p) and 54.50% (BGE-M3, +5.79%p). At 4B parameters, Qwen3-4B achieves 65.12% versus 59.57% (Jina-v4, +5.55%p). This consistent advantage across scales demonstrates that generative pretraining captures relational cues beyond discriminative training.

Disentangling Redundancy and Similarity. Passage similarity is more detrimental than redundancy. Using Qwen3-8B, we observe that despite Legal having the lowest redundancy (25.1%), it achieves only 67.16% Coverage@10 due to its high similarity (40.7%). In contrast, Finance—with 2.5× higher redundancy (63.2%) but lower similarity (35.1%)—reaches 72.92%. Patent demonstrates

Model	Size	Finance		Patent		Legal		HotpotQA	
		Coverage@10	Top@10	Coverage@10	Top@10	Coverage@10	Top@10	Coverage@10	Top@10
BM25	–	59.97	36.28	76.35	55.62	49.08	27.77	64.58	44.19
OpenAI-Large	–	67.67	42.95	81.96	61.46	61.17	37.36	92.78	86.48
E5-Large	0.56B	62.83	37.98	80.69	60.62	52.56	30.91	91.67	84.01
BGE-M3	0.56B	64.43	39.84	81.46	59.58	54.50	32.40	89.96	78.78
Qwen3	0.6B	67.62	43.80	81.49	60.21	60.29	35.37	**93.97**	**89.18**
Jina-v4	3.75B	65.10	40.31	80.09	59.58	59.57	34.55	92.61	85.04
Qwen3	4B	72.43	48.84	83.09	62.50	65.12	39.67	93.58	88.66
E5-Mistral	7B	69.69	44.81	81.67	62.71	56.86	34.05	93.53	88.66
Qwen3	8B	**72.92**	**48.84**	**84.05**	**63.12**	**67.16**	**41.49**	93.58	88.66

Table 3: Retrieval performance across domains. Dense retrievers outperform BM25, with smaller margins in high-overlap corpora. Performance improves with model size. LLM-based embeddings (Qwen3) consistently surpass pure embedding architectures (E5-Large, BGE-M3, Jina-v4) at comparable sizes.

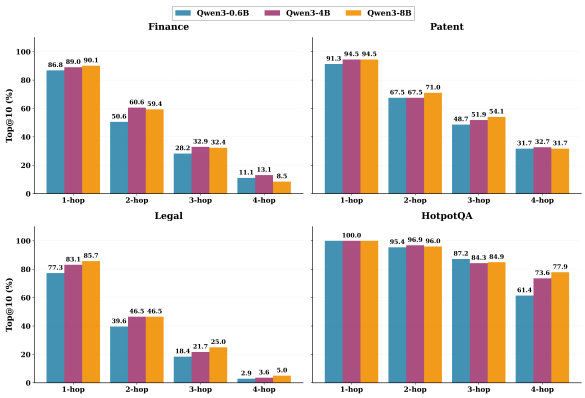

Figure 4: **Hop-wise Top@10 for Qwen3 models across sizes.** Scaling improves robustness in HotpotQA (0.6B: 61.4% → 8B: 77.9% at 4-hop), but high-similarity domains exhibit sharp degradation regardless of model size (Legal converges below 5% at 4-hop).

this pattern most clearly: with the lowest similarity (29.0%) and intermediate redundancy (49.7%), it attains the highest performance at 84.05%. High similarity forces retrievers to distinguish nearly identical passages, while redundancy provides multiple pathways to correct evidence.

4.3 Performance Degradation by Reasoning Depth

Figure 4 visualizes how performance degrades as hop depth increases from 1 to 4. Complete results appear in Appendix D.

Results. Degradation is driven by corpus properties rather than hop depth alone. Using Qwen3-8B, Top@10 drops from 90.1% to 8.5% (Finance), 85.7% to 5.0% (Legal), and 94.5% to 31.7% (Patent), while HotpotQA remains stable (100% to 77.9%). Notably, HotpotQA maintains 77.9% at 4-hop while Legal drops to 46.5% at only 2-hop, demonstrating that corpus characteristics dominate over reasoning depth.

Scaling improves robustness (+16.5%p from 0.6B to 8B in HotpotQA at 4-hop) but shows diminishing returns in high-similarity domains, where all Qwen3 variants converge below 15% (Finance) and 5% (Legal). Structural challenges impose fundamental limits beyond model capacity, suggesting architectural innovations—contrastive fine-tuning on similarity-aware objectives or explicit redundancy modeling—are necessary.

4.4 Effectiveness of SepRRF

This ablation evaluates whether separate criterion evaluation with rank fusion (SepRRF) improves quality over joint prompting or confidence-based aggregation. We compare three prompting strategies (Vanilla: holistic judgment; Combined: evaluate all criteria jointly; Separate: independent per-criterion evaluation) crossed with three aggregation methods (Base: arithmetic mean; MinMax: normalized mean; RRF: rank-based fusion). We sample 100 passages with the most sentences from HotpotQA's train and dev sets, and rank sentences within each passage using the five quality criteria from RARE's Valid Information Selection stage: validity, completeness, specificity, clarity, and questionability. Performance is measured by how highly each method ranks human-labeled evidence sentences, which we treat as implicit human preferences. Each configuration is run 5 times, with NDCG@3 as the primary metric and MRR and Top-1 accuracy in Appendix E.

Results. Figure 5 shows that SepRRF (Separate+RRF) achieves the highest performance: 46.3

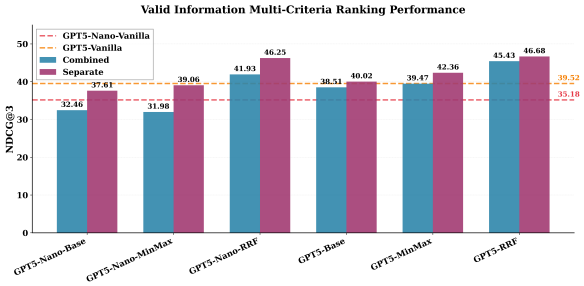

Figure 5: **SepRRF performance across prompting and aggregation strategies.** SepRRF achieves the highest NDCG@3 on both GPT-5 Nano and GPT-5. Separation helps more than explicit criteria alone, and RRF consistently outperforms confidence-based aggregation.

NDCG@3 on GPT-5 Nano and 46.7 on GPT-5. Three patterns emerge: **(1) Explicit criteria help**—Combined+RRF (41.9) outperforms Vanilla-Base (35.2) by 6.7%p. **(2) Separation helps more**—SepRRF (46.3) gains 4.3%p over Combined+RRF by enabling stable per-criterion judgments. **(3) RRF outperforms score-based aggregation**—under Separate prompting, RRF (46.3) exceeds MinMax (39.1) by 7.2%p, as it relies on stable ranks rather than unstable confidence magnitudes. Importantly, gains persist with stronger models: although GPT-5 outperforms GPT-5 Nano under Combined+RRF (45.4 vs 41.9), SepRRF still provides consistent improvement (46.7 vs 45.4), demonstrating that the benefit reflects fundamental advantages of decomposing complex multi-criterion reasoning.

5 Related Work

Open-Domain QA Benchmarks. Standard benchmarks (Kwiatkowski et al., 2019; Yang et al., 2018; Nguyen et al., 2016; Craswell et al., 2020) and evaluation suites (Thakur et al., 2021; Muennighoff et al., 2023) assume distinct documents with one-to-one answer-passage mappings. While this enables controlled evaluation, it abstracts away from real-world scenarios where identical information frequently recurs across multiple documents, making it difficult to assess retrieval robustness under redundancy.

Toward Realistic Benchmarks. Recent work has enhanced realism along various axes such as multilingual coverage (Zhang et al., 2022), long documents (Jin et al., 2023), and domain specificity (Petroni et al., 2021; Santhanam et al., 2021; Voorhees et al., 2021). However, these benchmarks do not address *high similarity and redundancy*—a defining characteristic of many enterprise corpora where the same information is repeatedly instantiated across documents with varying phrasings.

LLM-based Data Generation. Methods such as Promptagator (Dai and Callan, 2022), InPars (Bonifacio et al., 2022, 2023), and PAQ (Lewis et al., 2021) demonstrate automatic query generation from documents. However, these approaches primarily operate at the level of surface-form text generation or paraphrasing. In contrast, generating high-quality retrieval questions is a complex reasoning task that requires simultaneously considering multiple preference criteria while avoiding logical errors—an underexplored area requiring further investigation.

6 Conclusion

This work examined a fundamental limitation of current QA and RAG evaluation practices: most benchmarks assume distinct, non-overlapping documents, whereas real-world corpora—such as financial filings, legal codes, and patents—are characterized by substantial redundancy and inter-document similarity. This discrepancy can obscure true retrieval behavior, leading to over- or underestimation of model robustness.

We introduced **RARE**, a modular framework for constructing redundancy-aware retrieval benchmarks. RARE decomposes documents into atomic factual units to enable fine-grained redundancy tracking and employs **SepRRF**, a criterion-separated rank-fusion strategy, to stabilize LLM-based judgments without relying on poorly calibrated confidence scores. Applied to domain-specific corpora, RARE yields **RedQA**, an evaluation suite designed to reflect the structural properties of realistic RAG settings.

Empirical analyses suggest that retrieval performance degrades substantially under high-similarity and high-redundancy conditions, revealing a robustness gap that is largely invisible in conventional benchmarks. In particular, similarity among passages appears to pose a more persistent challenge than factual redundancy, limiting the effectiveness of both dense and sparse retrieval architectures. These findings indicate that standard benchmarks may overstate generalization to enterprise-scale deployments, underscoring the need for evaluation grounded in realistic corpus characteristics.

Limitations

RARE relies on LLM judges for both generation and verification, inheriting their biases and potential error modes. Redundancy detection uses a fixed similarity threshold ($\tau = 0.5$), which may require domain-specific tuning for optimal recall-precision trade-offs. While we validate SepRRF on sentence ranking in HotpotQA, broader generalization across diverse ranking tasks remains to be demonstrated.

Our evaluation focuses on retrieval quality; downstream generation quality in end-to-end RAG systems is not directly measured. As hop depth increases, some generated questions—although logically correct—tend to become artificially list-like, reflecting limitations of current generation approaches. Future work could address this by modeling atomic information connectivity through graph structures to enable more natural reasoning chains.

Ethics Statement

This work introduces a benchmark for evaluating retrieval systems in high-redundancy environments. The dataset construction uses publicly available documents (SEC filings, USPTO patents, U.S. Code, Wikipedia) and does not involve human participants or sensitive data.

LLM-based generation and filtering may inherit biases from the underlying models (GPT-5, GPT-5 Nano); we mitigate this through human evaluation (Section 3.1) showing 89.6%, 83.6% recall in detecting invalid content. RedQA is intended for research purposes to improve information retrieval systems, with potential societal benefits in enterprise search, legal research, and knowledge management. All data sources are used in accordance with their intended purposes and licensing terms.

References

Jing Bai and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2501.12345*.

Luiz Bonifacio and 1 others. 2022. Inpars: Data augmentation for information retrieval using large language models. In *Proceedings of the 2022 ACM SIGIR Conference on Research and Development in Information Retrieval*.

Luiz Bonifacio and 1 others. 2023. Inpars v2: Large language models as efficient data generators for information retrieval. In *Proceedings of the 2023 ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fuyu Chen and 1 others. 2024. Bge-m3: Multi-function, multi-lingual, multi-granularity text embedding model. *arXiv preprint arXiv:2402.03216*.

Gordon Cormack and Charles Clarke. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019)*.

Zhuyun Dai and Jamie Callan. 2022. Promptagator: Few-shot dense retrieval from large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Moritz Günther, Bo Hou, and 1 others. 2024. Jina embeddings v4: Universal text and image embeddings. *arXiv preprint arXiv:2408.01886*.

Gautier Izacard, Fabio Petroni, Lucas Hosseini, and 1 others. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Qiao Jin and 1 others. 2023. Mldr: A benchmark for multi-lingual document retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Vladimir Karpukhin and 1 others. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5357–5370. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Patrick Lewis and 1 others. 2021. Probably asked questions: Language model query generation for open-domain qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ruibo Liu and 1 others. 2023. Autoqagen: Automatic question generation for evaluating knowledge in llms. In *ACL*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers, and 1 others. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

OpenAI. 2024. text-embedding-3-large model card. `https://platform.openai.com/docs/guides/embeddings`.

Jiaxin Pan and 1 others. 2023. Evaluating question generation in llms: Parsing and consistency issues. *arXiv preprint arXiv:2312.10021*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. In *Foundations and Trends in Information Retrieval*.

Keshav Santhanam, Ming-Wei Chen, Barlas Oguz, Xilun Chen, Wen-tau Yih, and Patrick Lewis. 2021. Lotte: A benchmark for long-tail open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 625–639. Association for Computational Linguistics.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1333–1342.

Ellen M Voorhees. 2000. Overview of the trec 2000 question answering track. *TREC*.

Wei Wang and 1 others. 2022. Simlm and e5: Embedding models for information retrieval and beyond. In *arXiv preprint arXiv:2212.03533*.

Wei Wang and 1 others. 2024. E5-mistral: Improving long-text retrieval via mistral-based embedding models. *arXiv preprint arXiv:2405.08540*.

Zhe Wang and 1 others. 2023. Rankfusion: Robust rank aggregation for llm-based evaluation. *arXiv preprint arXiv:2311.11010*.

Zhiyuan Wu and 1 others. 2024. Rageval: Evaluating retrieval-augmented generation via evidence sensitivity. *arXiv preprint arXiv:2405.12219*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Xinyu Zhang and 1 others. 2022. Miracl: A multilingual retrieval dataset covering 18 languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kaiyu Zheng and 1 others. 2024. Legalbench-rag: A benchmark for retrieval-augmented legal reasoning. *arXiv preprint arXiv:2408.10343*.

A Dataset Construction Details

A.1 Domain Selection

Finance. SEC 10-K reports (2023–2024) from six technology companies: Apple, Microsoft, Amazon, Meta, Tesla, and NVIDIA. These reports exhibit high redundancy due to standardized disclosure requirements and repeated regulatory language.

Patents. USPTO documents (2023–2024) from five technical areas: H01M* (batteries and electrochemistry), H01L* (semiconductors and solid-state devices), H04W* (wireless communication networks), A61K* (pharmaceutical preparations), G06F* (computing and data processing). We selected 10 documents per area (50 total).

Legal. U.S. Code (2023) covering six titles: Title 1 (General Provisions), Title 9 (Arbitration), Title 17 (Copyrights), Title 35 (Patents), Title 40 (Public Buildings, Property, and Works), Title 41 (Public Contracts). Legal codes exhibit high similarity through cross-referencing and standardized legal language.

HotpotQA. Wikipedia passages extracted from the HotpotQA dataset, serving as a low-redundancy baseline where articles are topically distinct with minimal factual overlap.

A.2 Implementation Details

Hyperparameters. Multi-hop depth $M \in \{1, 2, 3, 4\}$; Atomic Information Pool Size in Question Generation $N = 100$; Question Candidates Size $S = 10$ per hop.

LLM Models. GPT-5 Nano for atomic extraction, filtering, and ranking stages. GPT-5 for question generation to encourage diverse candidate generation.

Chunking. Documents split into 512-token chunks using recursive text splitting with no overlap. Chunk boundaries respect sentence boundaries to maintain semantic coherence.

Embedding. OpenAI text-embedding-3-large (dimension=3072) for all similarity computations. Cosine similarity threshold $\tau = 0.5$ for redundancy detection candidate retrieval.

Top-k Selection. Top-3 atomic units per chunk selected after SepRRF ranking based on five quality criteria (validity, completeness, specificity, clarity, questionability).

Inference Hardware. NVIDIA L4 GPU with fp16 precision for all embedding computations. API-based LLM calls do not require local GPU resources.

Reproducibility. Fixed random seed (42) for all sampling operations.

A.3 Corpus Statistics

Table 4: Detailed corpus statistics across four domains.

Metric	Finance	Patent	Legal	HotpotQA
Chunks	3,281	3,339	4,090	3,218
Total atoms	42,825	39,282	24,815	26,093
Valid pass rate (%)	55.6	57.5	58.0	73.2
Top-3 atoms	6,040	6,875	7,717	8,232
Unique atoms	2,221	3,460	5,783	8,120
Redundancy (%)	63.2	49.7	25.1	1.4
Similarity (%)	35.1	29.0	40.7	8.8
Final samples	645	480	605	688

Valid pass rate indicates the proportion of extracted atomic units that pass the valid information filtering. HotpotQA's higher pass rate (73.2%) reflects cleaner Wikipedia text compared to real-world data.

Unique atoms are determined after redundancy tracking: atoms that do not have semantic equivalents in other chunks. The ratio of unique to top-3 atoms quantifies redundancy—Finance retains only 36.8% unique atoms (2,221/6,040), while HotpotQA retains 98.6% (8,120/8,232).

Table 5: Question filtering pass rates for Finance and Patent domains. All values in percentages. CI = Contextual Independence; AE = Answer Exclusion; IE = Information Equivalence; QC = Question Clarity; AS = Answerability.

Domain	Hop	CI	AE	IE	QC	AS	Success
Finance	1	70.8	74.4	97.2	99.0	98.4	50.7
	2	90.8	81.1	65.5	96.9	95.5	56.7
	3	90.2	79.0	75.1	99.0	94.0	56.7
	4	92.0	68.4	78.0	98.7	95.8	51.0
Patent	1	68.8	63.0	98.3	98.1	95.8	42.3
	2	80.1	66.9	65.2	98.4	93.8	38.0
	3	93.0	64.7	70.8	97.5	93.5	45.0
	4	88.1	51.8	74.8	97.5	93.8	34.7

Table 6: Question filtering pass rates for Legal and HotpotQA domains. All values in percentages. Abbreviations same as Table 5.

Domain	Hop	CI	AE	IE	QC	AS	Success
Legal	1	74.9	69.8	98.1	97.7	98.0	51.3
	2	86.9	75.0	69.7	98.3	94.9	53.0
	3	92.5	71.6	69.0	98.9	94.3	50.7
	4	92.4	63.4	77.2	97.6	92.4	46.7
HotpotQA	1	94.7	80.9	98.8	95.8	98.1	71.3
	2	90.7	72.0	88.9	95.3	98.7	58.3
	3	89.8	63.6	88.4	96.5	97.4	53.0
	4	86.2	56.9	85.9	92.2	98.6	46.7

A.4 Question Filtering Statistics

We generated 10 candidate questions per sample (300 samples × 4 hops × 4 domains × 10 paraphrased questions = 48,000 candidates total) and applied zero-tolerance logical filtering with five criteria. Tables 5 and 6 show pass rates by domain and hop depth.

Key Observations. Information Equivalence exhibits a characteristic U-shaped pattern: high at 1-hop (97–98%), dropping sharply at 2-hop (65–89%), then recovering at 3–4 hops (69–88%). This reflects LLM generation behavior: at 1-hop, matching question scope to a single atomic unit is straightforward; at 2-hop, models frequently under-constrain (requiring only one unit) or over-constrain (requiring additional unstated facts); at higher hops, more careful candidate construction improves alignment. HotpotQA maintains consistently higher pass rates (86–99%) than enterprise domains (65–98%), likely due to cleaner Wikipedia text and well-structured reasoning chains.

Answer Exclusion degrades monotonically with

hop depth across all domains, dropping from 63–81% at 1-hop to 52–69% at 4-hop. Multi-hop questions increasingly embed intermediate reasoning steps or partial answers within the question text itself, violating the requirement that questions must not reveal answer content.

Contextual Independence improves from 1-hop (69–95%) to higher hops (86–93%), except in HotpotQA which starts high and remains stable. At 1-hop, LLMs frequently generate meta-questions ("What section discusses X?") that reference document structure; this tendency decreases at higher hops where complex reasoning demands content-focused formulation.

Question Clarity and **Answerability** remain consistently high (92–99%) across all conditions, indicating that when LLMs generate questions, they typically produce well-formed and answerable queries—the primary challenge lies in precisely controlling question scope (Information Equivalence) and avoiding answer leakage (Answer Exclusion).

Overall Success Rates. Aggregating across all domains and hops: **2,418 questions retained from 4,800 generated (50.4% success rate)**. This reflects our conservative filtering approach where any single criterion failure results in rejection.

Domain-specific patterns reveal the impact of corpus characteristics on question generation difficulty. HotpotQA achieves the highest success rates (71.3% at 1-hop, declining to 46.7% at 4-hop), benefiting from cleaner Wikipedia text and well-structured reasoning chains. Enterprise domains show lower and more variable success: Finance maintains relatively stable rates (50.7–56.7%), Legal shows moderate performance (46.7–53.0%), while Patent exhibits the lowest success (34.7–45.0%), likely due to dense technical terminology and complex domain-specific constructs that challenge precise question formulation.

The 50.4% overall success rate—retaining half of generated candidates—reflects our design priority: dataset quality over quantity. By rejecting any candidate failing even one of five criteria, we ensure RedQA contains only logically sound, precisely scoped questions suitable for valid retrieval evaluation.

A.5 Cost Analysis

Question generation accounts for 69.7% of total cost ($460.07 / $660.10), driven by (i) generating 10 candidate questions per sample using GPT-5, and (ii) filtering each candidate against five strict logical criteria.

Table 7: Construction cost (USD) by domain and pipeline stage, based on OpenAI API pricing as of January 2025.

Stage	Finance	Patent	Legal	HotpotQA	Total
Atomic extraction	6.88	6.51	7.86	4.92	26.17
Valid selection	18.97	18.66	21.21	18.11	76.95
Embedding	0.07	0.01	0.06	0.04	0.18
Redundancy tracking	30.93	32.14	30.97	2.69	96.73
Question generation	110.03	112.31	110.56	127.17	460.07
Total per domain	166.88	169.63	170.66	152.93	**660.10**

Redundancy tracking is more expensive in high-redundancy domains (Finance: $30.93, Patent: $32.14) than in HotpotQA ($2.69) because the number of equivalence checks scales with redundancy: Finance's 6,040 top-3 atoms compress to only 2,221 unique atoms, requiring 3,819 LLM-based equivalence verifications.

Embedding costs are negligible ($0.18 total) because text-embedding-3-large is efficient and only requires one embedding per atomic unit.

Per-sample cost averages $0.28 (Finance), $0.35 (Patent), $0.28 (Legal), and $0.22 (HotpotQA), making RARE practical for constructing domain-specific benchmarks at scale.

B Atomic Information Examples from RedQA

Finance.

- *"As of December 31, 2023, maturities of sales-type lease receivables thereafter are $2 million."*
- *"Total goodwill from acquisitions in 2022 was 1,137 million."*
- *"Long-term debt was 52,623 million as of December 31, 2024."*
- *"Vision Pro is expected to be available in early calendar year 2024."*
- *"Apple manages its business primarily on a geographic basis."*
- *"We purchase memory from SK Hynix Inc."*
- *"The board will consider continued capital availability, market conditions, applicable laws and agreements when deciding whether to declare dividends."*

Legal.

- *"For competitive basis purchases, the notice must include the basis on which the selection will be made."*

- *"The Comptroller General shall conduct a 3-year study of applicants who file only in the United States."*
- *"The Commissioner of Patents shall maintain the application in a sealed condition and notify the applicant thereof."*
- *"An officer of the mechanical licensing collective may not simultaneously be an employee of any board member."*
- *"Each joint owner of a patent may make, use, offer to sell, or sell the patented invention."*
- *"Infrastructure projects include energy generation from fossil fuels."*

Patent.

- *"The cancer stem cell surface marker-binding agent can be labeled with a paramagnetic label."*
- *"Aluminum is a typical material for the positive current collector 110."*
- *"An aluminum nitride film can be used as the protection insulating layer."*
- *"The CMP composition contains 0.01 wt% to 1 wt% of at least one non-ionic surfactant."*
- *"The encoder unit is configured to receive content from a plurality of content providers."*
- *"CD140a is expressed on fetal human oligodendrocyte progenitor cells."*
- *"The transceiver is configured to receive a second command to reactivate."*

C Qualitative Examples

We provide representative multi-hop questions from RedQA to illustrate the reasoning complexity and evidence requirements at each hop depth.

C.1 1-Hop Questions

Example 1.1

Question: When Meta Platforms, Inc. prepares its consolidated financial statements, whose accounts are included in scope?

Answer: The accounts of Meta Platforms, Inc. itself, its subsidiaries in which Meta has a controlling financial interest, and any variable interest entities for which Meta is deemed the primary beneficiary.

Evidence: The consolidated financial statements include the accounts of Meta Platforms, Inc., its subsidiaries where Meta has controlling financial interests, and any variable interest entities for which Meta is deemed the primary beneficiary.

Example 1.2

Question: As part of a strategic review of large-model deployments, how long did it take to train a GPT with about 1.8 trillion parameters, and which hardware was used?

Answer: Duration: about 3–5 months; Hardware: 25,000 NVIDIA Ampere GPUs.

Evidence: Training a 1.8 trillion-parameter GPT model took about three to five months using 25,000 NVIDIA Ampere GPUs.

C.2 2-Hop Questions

Example 2.1

Question: Who is responsible for establishing and maintaining internal control over financial reporting, and which top executives formulate or supervise that framework?

Answer: Management is responsible for establishing and maintaining ICFR; the framework is designed or supervised by the Chief Executive Officer and the Chief Financial Officer.

Evidence: (1) Management is responsible for establishing and maintaining adequate internal control over financial reporting. (2) Internal control over financial reporting is a process designed by, or under the supervision of, our Chief Executive Officer and Chief Financial Officer.

Example 2.2

Question: What material is commonly used for the layer applied to the main portion of the anode current collector in a lithium-ion battery?

Answer: Graphitic carbon (carbon-based negative electrode material).

Evidence: (1) Graphitic carbon is commonly used as the negative-electrode material in lithium-ion batteries. (2) The main region of the anode current collector is loaded with a layer of negative electrode material.

C.3 3-Hop Questions

Example 3.1

Question: For a residential buyer of a Tesla energy storage system, what does the sale include, how long are the installation and components typically covered, and who provides the warranty if some components aren't made by Tesla?

Answer: The sale includes the installation of the energy storage system. The installation and the energy storage components are typically covered for 10–25 years. For components not manufactured by Tesla, Tesla generally passes through the applicable manufacturers' warranties (i.e., the third-party component warranties are provided by the original manufacturers).

Evidence: (1) Sales of energy storage systems to residential and small-scale commercial customers consist of the installation of the energy storage system. (2) The warranty on installation and components of Tesla's energy generation and storage systems lasts typically between 10 and 25 years. (3) For components not manufactured by us, we generally pass-through the applicable manufacturers' warranties.

Example 3.2

Question: For a quarterly market outlook, how could the Ethereum 2.0 upgrade affect demand for NVIDIA's latest graphics cards across its expanded customer base, given the possibility that higher resale of GPUs might dampen new-card purchases?

Answer: Ethereum 2.0 is likely to reduce Ethereum mining demand and increase aftermarket GPU supply, which could depress retail prices and reduce demand for NVIDIA's new GPUs. Because NVIDIA now serves broader markets beyond gaming, this negative effect on new-GPU demand would extend across its expanded customer base, not just gamers.

Evidence: (1) Changes to cryptocurrency standards and processes, including the Ethereum 2.0 merge, may create increased aftermarket sales of our GPUs. (2) NVIDIA's business has evolved from gaming products to broader markets. (3) Increased aftermarket sales of our GPUs could reduce demand for our new GPUs.

C.4 4-Hop Questions

Example 4.1

Question: Within NVIDIA's governance framework for a special meeting, who has the right to call the gathering, who presents annual information security reports to the Board and who receives them, who oversees internal audit, and who administers director orientation and continuing education?

Answer: Stockholders have the right to call the gathering; the Chief Security Officer and security team present annual information security reports to the Board (which receives them); the Board oversees internal audit; director orientation and continuing education are administered by the NCGC and General Counsel.

Evidence: (1) Stockholders can call a special meeting. (2) The NCGC and our General Counsel are responsible for new director orientation and for administering or approving eligible director continuing education programs. (3) The Board receives annual reports on information security matters from the Chief Security Officer and members of our security team. (4) The Board oversees the performance of the internal audit function.

Example 4.2

Question: In NVIDIA's compensation disclosures, which individuals are listed as non-CEO named executive officers, what form of equity dominates their target compensation, what is the maximum payout as a percentage of target for their target-based awards, and how does the CEO's equity setup compare?

Answer: Non-CEO NEOs: Colette M. Kress, Ajay K. Puri, Debora Shoquist, Timothy S. Teter. Their target pay is dominated by at-risk equity in the form of PSUs (60% PSUs, 40% RSUs). The maximum payout on target-based awards is 200% of target. The CEO's equity setup is different: Jensen Huang's target equity is 100% at-risk PSUs (split 50-50 between MY PSUs and SY PSUs).

Evidence: (1) For Fiscal 2023, the Compensation Committee (CC) decided that the largest portion of NEOs' total target pay would remain in the form of at-risk equity with performance-based vesting. (2) The maximum payout for NEOs under target-based awards is capped at 200% of target. (3) For Fiscal 2024, 2023, 2022,

Table 8: Finance: complete results by hop depth.

	Overall (645)				1-Hop (152)			
Model	Coverage@10	Top@10	NDCG@10	MRR	Coverage@10	Top@10	NDCG@10	MRR
BM25	60.0	36.3	48.6	36.2	82.9	82.9	64.6	58.7
OpenAI-Large	67.7	43.0	56.3	42.1	90.1	90.1	71.2	65.3
E5-Large	62.8	38.0	52.0	39.4	84.9	84.9	67.3	61.8
BGE-M3	64.4	39.8	52.6	39.3	83.6	83.6	64.7	58.7
Qwen3-0.6B	67.6	43.8	54.9	40.4	86.8	86.8	69.7	64.3
Jina-v4	65.1	40.3	55.2	42.2	86.2	86.2	74.4	70.7
Qwen3-4B	72.4	48.8	59.4	44.0	89.0	89.0	76.0	71.7
E5-Mistral	69.7	44.8	56.8	42.7	88.2	88.2	72.1	66.9
Qwen3-8B	72.9	48.4	60.4	44.8	90.1	90.1	75.4	70.7
	2-Hop (170)				**3-Hop (170)**			
BM25	60.9	37.1	47.8	35.7	52.0	20.6	43.0	28.5
OpenAI-Large	69.1	47.7	58.1	44.1	60.0	24.7	50.5	33.1
E5-Large	62.9	39.4	51.0	39.2	56.5	21.8	47.6	32.1
BGE-M3	67.1	44.7	54.3	41.9	55.5	22.9	46.4	29.7
Qwen3-0.6B	70.0	50.6	55.7	40.2	63.0	28.2	50.2	31.9
Jina-v4	66.5	45.3	55.8	43.0	58.0	23.5	47.2	30.8
Qwen3-4B	76.8	60.6	59.5	43.5	66.3	32.9	53.8	34.6
E5-Mistral	68.8	49.4	54.1	41.6	62.9	28.8	51.8	34.6
Qwen3-8B	77.9	59.4	61.2	45.3	66.1	32.4	55.6	36.1
	4-Hop (153)							
BM25	45.1	6.5	39.6	22.8				
OpenAI-Large	52.3	11.1	45.8	26.8				
E5-Large	51.0	7.8	42.7	25.6				
BGE-M3	52.5	9.8	45.5	27.6				
Qwen3-0.6B	53.3	11.1	45.2	26.4				
Jina-v4	51.0	7.8	44.2	25.7				
Qwen3-4B	57.5	13.1	48.7	27.5				
E5-Mistral	59.8	14.4	50.1	29.0				
Qwen3-8B	57.8	8.5	49.8	28.2				

Table 9: Patent: complete results by hop depth.

	Overall (480)				1-Hop (127)			
Model	Coverage@10	Top@10	NDCG@10	MRR	Coverage@10	Top@10	NDCG@10	MRR
BM25	76.4	55.6	63.2	48.8	88.2	88.2	71.2	65.7
OpenAI-Large	82.0	61.5	66.5	49.8	92.1	92.1	73.2	67.0
E5-Large	80.7	60.6	64.9	48.2	90.6	90.6	69.3	62.4
BGE-M3	81.5	59.6	67.4	51.0	91.3	91.3	75.1	69.0
Qwen3-0.6B	81.5	60.2	68.0	51.5	91.3	91.3	73.0	67.0
Jina-v4	80.1	59.6	66.6	51.0	89.0	89.0	70.2	64.3
Qwen3-4B	83.1	62.5	70.8	55.1	94.5	94.5	76.9	71.3
E5-Mistral	81.7	62.7	65.7	48.6	92.1	92.1	70.6	63.6
Qwen3-8B	84.1	63.1	71.4	54.6	94.5	94.5	77.5	72.1
	2-Hop (114)				**3-Hop (135)**			
BM25	77.2	61.4	61.6	49.9	74.8	44.4	63.2	43.0
OpenAI-Large	84.2	70.2	65.9	49.2	79.0	50.4	66.5	43.9
E5-Large	77.2	64.0	60.6	45.5	78.5	52.6	66.1	43.6
BGE-M3	82.5	65.8	63.9	48.2	78.5	49.6	67.0	44.5
Qwen3-0.6B	81.6	67.5	65.0	50.9	80.2	48.7	69.0	45.6
Jina-v4	78.5	63.2	62.3	48.3	79.8	51.1	69.3	47.9
Qwen3-4B	83.3	67.5	68.7	54.2	80.6	51.9	70.6	49.9
E5-Mistral	81.1	67.5	63.3	46.5	80.7	57.0	66.9	44.2
Qwen3-8B	85.1	71.1	71.0	55.3	80.5	54.1	70.0	47.2
	4-Hop (104)							
BM25	63.0	24.0	55.1	35.4				
OpenAI-Large	70.9	28.9	59.9	36.9				
E5-Large	70.9	31.7	63.2	42.1				
BGE-M3	73.1	27.9	63.0	40.7				
Qwen3-0.6B	72.6	31.7	66.0	43.0				
Jina-v4	73.1	26.9	62.7	40.7				
Qwen3-4B	74.5	32.7	66.6	42.7				
E5-Mistral	73.1	28.9	62.8	38.8				
Qwen3-8B	72.6	31.7	65.7	43.0				

and 2021, our non-CEO NEOs were Colette M. Kress, Ajay K. Puri, Debora Shoquist, and Timothy S. Teter. (4) Mr. Huang, the CEO, has 100% of his equity grants designated as at-risk and performance-based.

C.5 Observations

As hop depth increases, questions require synthesizing information from progressively more dispersed evidence passages. At 4-hop, retrievers must aggregate facts across governance structures, reporting chains, regulatory contexts, and specific role definitions—all while navigating high document redundancy that introduces numerous near-duplicate distractors. This complexity mirrors real-world RAG deployments where answers rarely reside in a single passage.

D Complete Experimental Results

D.1 Main Results: All Metrics

Tables 8 through 11 present complete results (Coverage@10, Top@10, NDCG@10, MRR) across all models, domains, and hop depths.

E SepRRF Ablation Study

We evaluate whether separate criterion evaluation combined with rank fusion (SepRRF) improves ranking quality over joint prompting or confidence-based aggregation. This ablation isolates the contributions of (1) prompting strategy and (2) aggregation method.

Table 10: Legal: complete results by hop depth.

	Overall (605)				1-Hop (154)			
Model	Coverage@10	Top@10	NDCG@10	MRR	Coverage@10	Top@10	NDCG@10	MRR
BM25	49.1	27.8	40.3	29.9	68.8	68.8	56.9	53.1
OpenAI-Large	61.2	37.4	48.9	35.2	79.2	79.2	60.2	54.2
E5-Large	52.6	30.9	40.8	29.0	64.9	64.9	48.7	43.6
BGE-M3	54.5	32.4	42.9	30.7	70.1	70.1	51.7	45.9
Qwen3-0.6B	60.3	35.4	46.6	33.1	77.3	77.3	56.0	49.4
Jina-v4	59.6	34.6	47.4	34.0	76.6	76.6	57.4	51.3
Qwen3-4B	65.1	39.7	53.8	39.1	83.1	83.1	64.2	58.1
E5-Mistral	56.9	34.1	43.3	31.0	71.4	71.4	52.8	46.9
Qwen3-8B	67.2	41.5	57.1	42.5	85.7	85.7	70.4	65.6
	2-Hop (159)				**3-Hop (152)**			
BM25	51.9	29.6	39.7	29.1	38.6	8.6	32.9	20.5
OpenAI-Large	64.5	44.7	51.0	36.8	56.8	19.1	47.0	29.2
E5-Large	57.6	36.5	43.5	31.3	49.3	17.8	39.9	24.4
BGE-M3	59.8	40.9	45.5	32.8	48.7	14.5	42.8	27.6
Qwen3-0.6B	62.9	39.6	46.3	33.1	55.3	18.4	46.6	29.4
Jina-v4	59.4	39.0	45.9	32.9	58.3	18.4	48.9	31.4
Qwen3-4B	66.4	46.5	53.5	40.1	59.9	21.7	53.3	33.1
E5-Mistral	62.6	40.3	47.8	35.3	53.3	17.1	41.0	24.5
Qwen3-8B	68.2	46.5	56.5	43.7	64.0	25.0	55.3	34.9
	4-Hop (140)							
BM25	35.5	1.4	30.6	15.6				
OpenAI-Large	42.3	6.4	36.3	19.3				
E5-Large	42.3	3.6	30.2	15.4				
BGE-M3	37.7	0.7	30.2	14.9				
Qwen3-0.6B	46.4	2.9	36.9	19.1				
Jina-v4	42.9	1.4	34.5	15.1				
Qwen3-4B	49.6	3.6	44.5	23.7				
E5-Mistral	38.2	4.3	30.5	15.6				
Qwen3-8B	57.3	5.0	46.6	25.7				

E.1 Experimental Setup

Task. We rank evidence sentences from HotpotQA based on five quality criteria used in RARE's Valid Information Selection stage: validity, completeness, specificity, clarity, and questionability. Each method ranks 100 candidate sentences, and we measure NDCG@3, MRR, and Top-1 accuracy against human judgments. Since HotpotQA passages are shorter than those typically found in industry settings, we combined the train and dev sets and selected the top 100 passages by token length, resulting in an average of 350 tokens—closely matching RedQA's average passage length of 356.95 tokens.

Prompting Strategies.

- **Vanilla**: Single holistic prompt asking to rank sentences by overall quality without explicit cri-

Table 11: HotpotQA: complete results by hop depth.

Model	Overall (688)				1-Hop (145)			
	Coverage@10	Top@10	NDCG@10	MRR	Coverage@10	Top@10	NDCG@10	MRR
BM25	64.6	44.2	61.8	48.6	86.9	86.9	80.5	78.4
OpenAI-Large	92.8	86.5	92.2	70.2	100	100	99.8	99.8
E5-Large	91.7	84.0	90.8	69.3	100	100	99.1	98.8
BGE-M3	90.0	78.8	88.8	68.0	100	100	99.1	98.8
Qwen3-0.6B	94.0	89.2	92.9	70.5	100	100	98.8	98.4
Jina-v4	92.6	85.0	91.3	69.6	100	100	99.5	99.3
Qwen3-4B	93.6	88.7	92.6	70.3	100	100	99.0	98.4
E5-Mistral	93.5	88.7	92.3	70.2	100	100	99.4	99.2
Qwen3-8B	93.6	88.7	92.6	70.3	100	100	99.0	98.4
	2-Hop (175)				**3-Hop (159)**			
BM25	70.9	51.4	67.5	50.6	49.3	13.8	48.0	29.3
OpenAI-Large	92.0	92.0	93.9	70.6	89.5	83.0	88.9	54.3
E5-Large	93.7	93.7	93.5	70.6	88.1	76.1	87.4	52.7
BGE-M3	94.0	94.3	93.6	70.1	85.9	75.5	84.5	51.6
Qwen3-0.6B	95.4	95.4	95.2	71.5	87.2	75.5	85.2	51.7
Jina-v4	93.7	90.3	93.1	69.9	88.7	79.9	87.6	53.4
Qwen3-4B	96.9	96.9	95.0	71.5	89.9	84.3	89.6	54.5
E5-Mistral	96.0	94.3	93.8	70.8	90.6	85.5	89.0	54.0
Qwen3-8B	96.9	96.9	95.0	71.5	90.2	84.9	89.0	54.1
	4-Hop (209)							
BM25	40.0	4.3	41.8	22.6				
OpenAI-Large	81.3	62.9	82.0	42.8				
E5-Large	81.9	61.4	83.1	42.4				
BGE-M3	75.7	47.9	75.9	39.9				
Qwen3-0.6B	81.4	61.4	83.1	42.5				
Jina-v4	81.4	63.4	82.5	42.9				
Qwen3-4B	85.0	73.6	84.4	44.7				
E5-Mistral	83.9	67.9	83.6	43.5				
Qwen3-8B	87.1	77.9	87.6	45.1				

teria enumeration.

- **Combined**: Single prompt listing all five criteria and asking the LLM to jointly evaluate and rank sentences considering all criteria simultaneously.
- **Separate**: Five independent prompts, each evaluating one criterion. Each prompt produces a per-criterion ranking without considering other criteria.

Aggregation Methods.

- **Base**: Arithmetic mean of raw LLM confidence scores across criteria. For Vanilla (single criterion), this is the raw score.
- **MinMax**: Min-max normalization of scores to $[0, 1]$ before averaging, addressing scale mismatches between criteria.
- **RRF** (Reciprocal Rank Fusion): Rank-based fusion $s(x) = \sum_{i=1}^{N} \frac{1}{\text{rank}_i(x)}$, discarding confidence magnitudes entirely and relying only on ordinal preferences.

We run each configuration 5 times and report mean values.

E.2 Complete Ablation Results

Tables 12 and 13 present results for GPT-5 Nano and GPT-5 (5-run average).

E.3 Stability Analysis

Beyond mean performance, stability across runs is critical for production deployment. Table 14 shows coefficient of variation (CV = std/mean × 100%) across 5 runs.

Separate Prompting Reduces Variance. Comparing prompting strategies within the same aggregation method reveals consistent patterns:

Table 12: SepRRF ablation on GPT-5 Nano (5-run avg). Separate+RRF achieves substantial gains over all baselines.

Prompt	Agg	NDCG@3	MRR	Top-1
Vanilla	Base	35.18	40.10	22.00
Combined	Base	32.46	37.06	16.20
Combined	MinMax	31.98	37.55	17.60
Combined	RRF	41.93	45.20	26.20
Separate	Base	37.61	42.25	23.60
Separate	MinMax	39.06	43.48	25.00
Separate	RRF	**46.25**	**49.39**	**33.00**

Table 13: SepRRF ablation on GPT-5 (5-run avg). Gains persist with stronger models, confirming fundamental advantages of decomposed reasoning.

Prompt	Agg	NDCG@3	MRR	Top-1
Vanilla	Base	39.52	44.68	27.80
Combined	Base	38.51	42.91	24.00
Combined	MinMax	39.47	43.39	23.60
Combined	RRF	45.43	48.95	32.20
Separate	Base	40.02	45.22	28.40
Separate	MinMax	42.36	46.26	28.20
Separate	RRF	**46.68**	**49.71**	**33.20**

Separate vs. Combined (Base): Separate reduces CV from 5.8% to 4.1% (Nano) and from 4.5% to 3.1% (GPT-5) for NDCG@3. Independent criterion evaluation produces more stable rankings because variations in one criterion do not cascade to affect judgments of other criteria.

Separate vs. Vanilla: Separate+Base (4.1% CV) shows lower variance than Vanilla+Base (5.2%) on Nano, suggesting that explicit criterion decomposition—even without rank fusion—improves consistency by providing structured evaluation scaffolding.

RRF Dramatically Stabilizes Combined Prompting. The most striking pattern emerges in Combined prompting: RRF reduces NDCG@3 CV from 6.6% (MinMax) to 1.7%—a 3.9× improvement. This shows that rank-based fusion acts as a powerful variance-reduction mechanism even when individual judgments are unstable. By aggregating ordinal preferences from multiple criteria rather than relying on single-criterion cardinal scores, RRF introduces implicit ensembling that smooths out noise.

Model-Specific Stability Patterns. GPT-5 Nano shows a clear stability ranking: Separate+RRF

Table 14: Cross-run variability (coefficient of variation in %) across 5 runs. Lower values indicate more consistent behavior.

		GPT-5 Nano		GPT-5	
Prompt	**Agg**	**NDCG@3**	**MRR**	**NDCG@3**	**MRR**
Vanilla	Base	5.2	5.1	2.7	2.0
Combined	Base	5.8	4.0	4.5	3.0
Combined	MinMax	6.6	5.4	3.7	2.3
Combined	RRF	1.7	1.4	4.3	2.8
Separate	Base	4.1	2.8	3.1	1.2
Separate	MinMax	3.3	2.8	**0.9**	**0.5**
Separate	RRF	**3.2**	**1.7**	1.7	0.8

(3.2%) < Separate+MinMax (3.3%) < Separate+Base (4.1%), with all three outperforming Combined+MinMax (6.6%). The minimal gap between RRF and MinMax (0.1 percentage point) reflects the fundamental stability of separate prompting, where aggregation method has less impact.

GPT-5 exhibits different behavior: Separate+MinMax achieves the lowest CV (0.9% NDCG@3, 0.5% MRR), outperforming even Separate+RRF (1.7%, 0.8%). This suggests that stronger models produce more calibrated confidence scores where normalization-based aggregation becomes viable. However, RRF maintains the highest absolute performance (46.68 vs. 42.36 NDCG@3), demonstrating a stability-performance tradeoff.

MRR vs. NDCG@3 Stability. MRR consistently shows lower CV than NDCG@3 across all configurations (e.g., Separate+RRF: 1.7% vs. 3.2% on Nano). This occurs because MRR focuses on the rank of the single highest-quality item, which is typically more stable than the aggregate quality of the top-3 items captured by NDCG@3.

E.4 Implications for RARE

These results directly motivate RARE's design choices:

(1) Decompose complex judgments. Rather than asking LLMs to rank atomic information or questions by considering all quality criteria simultaneously, RARE evaluates each criterion through independent prompts. This reduces per-call complexity and improves both reliability and consistency.

(2) Fuse via ranks, not scores. RARE aggregates criterion-specific rankings using RRF rather than averaging confidence scores. While Separate+MinMax achieves comparable or even superior stability on stronger models (0.9% CV on GPT-5), RRF maintains substantially higher absolute performance (46.68 vs. 42.36 NDCG@3), making it the preferred choice when optimizing for quality rather than variance minimization alone.

(3) Prioritize stability for benchmark construction. In dataset generation, consistency is as important as peak performance. SepRRF's low cross-run variance (1.7–3.2% CV for NDCG@3) ensures that dataset quality remains stable across construction runs, avoiding the need for extensive manual post-hoc filtering that would arise from unstable generation. The 6.6% CV of Combined+MinMax would introduce unacceptable variability, potentially requiring multiple regeneration attempts to achieve acceptable quality.

By adopting SepRRF throughout the pipeline—for both atomic information ranking (5 criteria) and question ranking (4 criteria)—RARE achieves measurably higher quality and greater consistency than alternative multi-criterion judgment strategies.

F Prompt Design and Examples

This section describes the prompt design for each stage of the RARE pipeline with representative examples of good and bad outputs.

F.1 Atomic Information Extraction

Prompt Design. The prompt instructs the LLM to extract atomic information units that will serve as ground truth answers for RAG dataset generation. Each unit must satisfy three critical criteria: (1) **Atomicity**—contains exactly one indivisible factual claim that cannot be meaningfully split, (2) **Validity**—provides substantively useful knowledge that supports real-world queries rather than trivial metadata, and (3) **Unambiguity**—completely self-contained with all entities explicitly named and no vague references. The prompt requires step-by-step reasoning for each criterion and outputs structured JSON.

Good Examples.

- *"Tesla's headquarters is located in Austin, Texas."* (single location fact, unambiguous)
- *"The United States Declaration of Independence was adopted in 1776."* (one historical fact, self-contained)
- *"The U.S. Supreme Court has nine justices."* (specific structural fact)

Bad Examples.

- *"Tesla's headquarters is in Austin, and Apple's headquarters is in Cupertino."* (violates atomicity—two facts)
- *"The study shows that healthcare costs are rising."* (violates unambiguity—which study?)
- *"What is the page of the document?"* (violates validity—trivial metadata)

F.2 Valid Information Filtering

Prompt Design. The prompt detects information completeness errors in atomic units. It filters out units that lack essential components, contain incomplete facts, or provide insufficient information for RAG applications. The prompt identifies five critical error patterns: (1) Missing essential elements (WHO/WHAT/WHEN/WHERE/HOW), (2) Fragmented information, (3) Contextually insufficient information, (4) Utility-deficient information (trivial facts, circular definitions), and (5) Meta information (document metadata, URLs, structural information). The output is a binary pass/fail decision with reasoning.

Good Examples (Pass).

- *"Albert Einstein discovered the theory of relativity in 1905."* (complete: who, what, when)
- *"Disney announced its acquisition of 21st Century Fox for $71.3 billion in December 2017."* (all essential details)
- *"Amazon's quarterly sales increased by 15% to $125.5 billion in Q3 2023 compared to Q3 2022."* (specific quantitative context)

Bad Examples (Fail).

- *"Discovered the theory of relativity."* (missing who, when)
- *"The merger was announced."* (missing which companies, when, financial terms)
- *"The document title is tesla_2023_annual_report.pdf."* (meta information with no substantive value)

F.3 Valid Information Multi-Criteria Ranking

Prompt Design. Five separate prompts rank atomic units across quality dimensions, each evaluated independently before combining with SepRRF. The prompts instruct the LLM to provide numerical scores (0-1) with step-by-step reasoning.

F.3.1 Criterion 1: Validity

Purpose. Evaluates retrieval usefulness for RAG systems—whether information meaningfully supports answering user queries. Filters out overly abstract content, trivial notes, structural markup, or metadata-only fragments.

Good Examples.

- *"The Great Wall of China is over 21,000 kilometers long."*
- *"Water boils at 100 degrees Celsius at standard atmospheric pressure."*
- *"Shakespeare wrote the play Hamlet around 1600."*

Bad Examples.

- *"Books contain pages with text."* (trivial common knowledge)
- *"The sky appears during daytime."* (obvious, low value)
- *"People use phones to make calls."* (not meaningfully useful)

F.3.2 Criterion 2: Completeness

Purpose. Assesses whether information is self-contained and understandable without additional document context. Someone unfamiliar with the source should comprehend it independently.

Good Examples.

- *"The Amazon River is the second longest river in the world."*
- *"Photosynthesis is the process by which plants convert sunlight into chemical energy."*
- *"The capital city of Japan is Tokyo."*

Bad Examples.

- *"It is the second longest river in the world."* (missing subject)
- *"This process is the way plants convert sunlight into chemical energy."* (vague reference)
- *"The capital city there is Tokyo."* (ambiguous location)

F.3.3 Criterion 3: Specificity

Purpose. Evaluates degree of concrete detail—presence of specific numbers, dates, names, entities, procedures, or conditions rather than vague generalities.

Good Examples.

- *"Mount Everest has a height of about 8,849 meters."*
- *"The Constitution of the United States was ratified on September 17, 1787."*
- *"DNA is composed of four bases: adenine, thymine, guanine, and cytosine."*

Bad Examples.

- *"Mount Everest is very tall."* (no specific measurement)
- *"The U.S. Constitution was created a long time ago."* (vague temporal reference)
- *"DNA is made up of several components."* (lacks enumeration)

F.3.4 Criterion 4: Clarity

Purpose. Ensures information is unambiguous with clear, unidirectional claims. Avoids polysemy, vague pronouns, or statements whose impact remains open to multiple interpretations.

Good Examples.

- *"Intense exercise increases the risk of accidents."*
- *"A vegetarian diet reduces the risk of cardiovascular disease."*
- *"Internet addiction lowers academic achievement."*

Bad Examples.

- *"Intense exercise has a major impact."* (unclear direction)
- *"A vegetarian diet brings many changes to health."* (ambiguous effect)
- *"The internet has an important effect on society."* (unspecified nature)

F.3.5 Criterion 5: Questionability

Purpose. Assesses whether information naturally forms the basis of a well-formed, specific question with a clear, factual, bounded answer suitable for retrieval-based QA.

Good Examples.

- *"Isaac Newton published Principia in 1687."* → When did Newton publish *Principia*?
- *"Antarctica is the coldest continent on Earth."* → Which is the coldest continent?
- *"The Berlin Wall collapsed in 1989."* → When did the Berlin Wall collapse?

Bad Examples.

- *"Principia had a major influence on multiple fields."* (too broad, no single answer)
- *"Antarctic exploration holds important significance in human history."* (interpretive, open-ended)
- *"The collapse of the Berlin Wall was a major historical event."* (not retrieval-friendly)

F.4 Redundancy Detection

Prompt Design. The prompt determines whether multiple information units convey identical core facts regardless of surface-level differences in expression. It evaluates three types of equivalence: (1) **Semantic Equivalence**—same core meaning and conceptual content despite different wording, (2) **Logical Equivalence**—identical logical structure, reasoning patterns, and cause-effect relationships, and (3) **Factual Equivalence**—same entities, properties, relationships, and quantitative/qualitative aspects. The output is list of binary: REDUNDANT (information units are equivalent) or UNIQUE (distinct content).

REDUNDANT

- *Target: "The company's revenue increased by 15% in Q3 2023."*
  Comparison: "In the third quarter of 2023, company earnings grew fifteen percent."
  (Same core meaning, logical relationship, entity, time period, metric, and value)
- *Target: "Water boils at 100 degrees Celsius at sea level."*
  Comparison: "At standard atmospheric pressure, H_2O reaches boiling point at 100°C."
  (Equivalent despite different terminology)

- *Target: "Shakespeare wrote Hamlet around 1600."*
  Comparison: "The play Hamlet was authored by William Shakespeare circa 1600."
  (Identical factual content, different expression)

UNIQUE

- *Target: "The company's revenue increased by 15% in Q3 2023."*
  Comparison: "The company's expenses decreased by 10% in Q3 2023."
  (Different concepts: revenue vs. expenses; opposite directions: increase vs. decrease)

- *Target: "Apple's iPhone sales grew in North America."*
  Comparison: "Apple's iPad sales grew in North America."
  (Different products despite same company and region)

- *Target: "The patent was filed in 2023."*
  Comparison: "The patent was granted in 2023."
  (Different events: filing vs. granting)

F.5 Question Generation

Prompt Design. The prompt operates in two steps: (1) **Information Selection**—from a pool of atomic units, select items with highest connectivity potential for natural multi-hop reasoning, and (2) **Question Generation**—generate multiple unique questions requiring all selected units. The generation follows a two-level structure: **Level 1: Mandatory Logical Consistency** (zero-tolerance—all 4 criteria must pass) and **Level 2: Preferred Quality Dimensions** (for human preference ranking). Each generated question includes explicit reasoning checks for all 8 dimensions.

F.5.1 Information Selection Criteria

The prompt selects units based on: (1) Semantic connectivity—natural logical relationships or causal chains, (2) Complementary information—different pieces of a larger puzzle, (3) Reasoning chain potential—elegant paths vs. parallel facts, and (4) Natural integration—how humans naturally think about topics. Avoids completely independent facts, redundant information, or single-hop answerable items.

F.5.2 Level 1: Mandatory Logical Consistency

These four criteria mirror the Question Logical Filtering criteria (Section F.6) and must all pass with zero tolerance:

1. Contextual Independence

- *Bad: "According to the table above, which environmental policy is most effective?"*
- *Good: "Which environmental policy in American Health Care Act is most effective?"*

2. Answer Exclusion

- *Bad: "Marie Curie worked at the University of Paris. When was the University of Paris established?"*
- *Good: "What year was the university where Marie Curie worked established?"*

3. Information Equivalence

- *Bad (Overflow): "What university did Marie Curie work at?"* (only needs one unit)
- *Bad (Underflow): "How many students does the university where Marie Curie worked have?"* (requires external info)
- *Good: "What year was the university where Marie Curie worked established?"* (requires exactly both units)

4. Question Ambiguity

- *Bad: "Where is our headquarter located?"* (ambiguous pronoun)
- *Good: "Where is Tesla's headquarters located?"* (clear entity reference)

F.5.3 Level 2: Preferred Quality Dimensions

These four dimensions mirror the Question Multi-Criteria Ranking criteria (Section F.7):

1. Connectivity

- *Bad: "In what year was Apple founded, and in what year was the iPhone released?"* (parallel listing)
- *Good: "How many years did it take Apple to release its first iPhone after its founding?"* (unified calculation)

2. Fluency

- *Bad: "What is the temporal differential calculation when subtracting the institutional establishment year from 1900?"*
- *Good: "In what year did the female scientist who won the Nobel Prize twice receive the Nobel Prize in Chemistry?"*

3. Essentiality

- *Bad: "What is the exact chronological period during which the highly distinguished Nobel laureate Marie Curie conducted her groundbreaking research at the prestigious University of Paris?"*
- *Good: "When did Marie Curie work at the University of Paris?"*

4. Validity

- *Bad: "What is the exact number of letters in the name of the university where Marie Curie worked?"*
- *Good: "What is considered Marie Curie's greatest achievement?"*

F.6 Question Logical Filtering

Prompt Design. Five filtering prompts implement zero-tolerance logical criteria. Each criterion must be satisfied independently; failure on any single criterion results in question rejection. The prompts enforce strict logical consistency to ensure questions are well-formed for RAG evaluation.

F.6.1 Criterion 1: Contextual Independence

Purpose. Detects questions that improperly assume or reference external document structures, sentence numbering, or meta-textual elements. Questions must be self-contained and understandable without knowledge of their presentation format.

Good Examples (Pass).

- *"What are the primary effects of climate change on coastal ecosystems?"*
- *"Which renewable energy sources have the highest efficiency ratings?"*
- *"What methods are most effective for carbon emission reduction?"*

Bad Examples (Fail).

- *"What climate change effects are mentioned in the second sentence?"* (sentence numbering reference)
- *"Based on the document provided, what are the primary renewable energy sources?"* (document reference)
- *"According to the table above, which environmental policy is most effective?"* (structural reference)

F.6.2 Criterion 2: Answer Exclusion

Purpose. Identifies questions that embed partial or complete answers within the question formulation, eliminating genuine multi-hop reasoning requirements. Questions must not contain pre-provided intermediate reasoning steps.

Good Examples (Pass).

- *"What year was the university where Marie Curie worked established?"* (requires two-hop: identify university, then find establishment date)
- *"What year was the founder of Microsoft born?"* (requires: identify founder, then find birth year)
- *"Which acquisition did Tesla's CEO complete in 2022?"* (requires: identify CEO, then find acquisition)

Bad Examples (Fail).

- *"Marie Curie worked at the University of Paris. When was the University of Paris established?"* (embeds first hop answer)
- *"The founder of Microsoft, Bill Gates, was born in which year?"* (reveals founder identity)
- *"Tesla's CEO Elon Musk made which acquisition in 2022?"* (provides CEO name)

F.6.3 Criterion 3: Information Equivalence

Purpose. Enforces the Perfect Information Equivalence Equation: questions must be fully answerable using only the provided atomic units and must require all units without redundancy. Detects two error patterns: (1) **Information Overflow**—questions solvable with only a subset of provided units, and (2) **Information Underflow**—questions requiring external knowledge beyond provided units.

Good Examples (Pass).

- *Provided: "Microsoft was founded by Bill Gates," "Bill Gates was born in 1955," "Microsoft headquarters is in Redmond, Washington"*
  Question: "Where is Microsoft's headquarters located, and when was the founder born?"
  (Requires all three units; no overflow or underflow)

- *Provided: "Apple Inc. was founded in 1976," "Steve Jobs co-founded Apple with Steve Wozniak," "Apple's HQ is in Cupertino"*
  Question: "When was the company co-founded by Steve Jobs and Steve Wozniak established, and where is it headquartered?"
  (Information Required = Information Provided exactly)

Bad Examples (Fail).

- *Provided: "Microsoft was founded by Bill Gates," "Bill Gates was born in 1955," "Microsoft HQ is in Redmond"*
  Question: "Who founded Microsoft?"
  (Information Overflow—only first unit needed; units 2 and 3 redundant)

- *Provided: "Tesla released its first EV in 2008," "Elon Musk is CEO of Tesla"*
  Question: "What was Tesla's stock price when it went public?"
  (Information Underflow—IPO information not provided)

F.6.4 Criterion 4: Question Clarity

Purpose. Detects questions with vague referential terms, underspecified comparisons, or context-dependent queries. All referential terms must have clear antecedents, and comparisons must include explicit reference points.

Good Examples (Pass).

- *"Who directed the 1994 film 'The Lion King'?"* (specific movie reference)

- *"Between Steven Spielberg and Martin Scorsese, who has directed films for more years?"* (explicit comparison basis)

- *"What factors influenced Apple's decision to discontinue the iPod Classic in 2014?"* (clear entity and event)

Bad Examples (Fail).

- *"That movie was directed by whom?"* (vague reference—which movie?)

- *"Who worked in the field longer?"* (underspecified—which field? compared to whom?)

- *"When was the company more profitable?"* (ambiguous—which company? compared to when?)

F.6.5 Criterion 5: Answerability

Purpose. Determines whether questions can be completely answered using only the provided document chunks. Questions must have all necessary information present without requiring external knowledge or critical information gaps.

Good Examples (Pass).

- *Chunks contain: revenue figures, growth percentages, time periods*
  Question: "What was the company's revenue growth rate in Q3 2023?"
  (All information present for complete answer)

- *Chunks contain: founding date, founder names, company location*
  Question: "When and where was the company founded, and by whom?"
  (Complete information available across chunks)

Bad Examples (Fail).

- *Chunks contain: product launch date, product features*
  Question: "What was the market share after the product launch?"
  (Market share data missing from chunks)

- *Chunks contain: company name, CEO name*
  Question: "What was the CEO's previous work experience before joining?"
  (Previous employment history not in chunks)

F.7 Question Multi-Criteria Ranking

Prompt Design. Four ranking prompts evaluate candidate questions across quality dimensions. Each prompt provides numerical scores (0-1) with reasoning. Rankings are combined using SepRRF to select the final question.

F.7.1 Criterion 1: Connectivity

Purpose. Measures logical flow and integration elegance in multi-hop reasoning. Evaluates how naturally the question connects multiple information sources through four dimensions: (1) Logical flow smoothness—natural transitions between reasoning steps, (2) Answer synthesis quality—coherent unified result vs. parallel enumeration, (3) Reasoning chain elegance—intellectually satisfying path, and (4) Information integration depth—meaningful semantic connections.

Good Examples (High Connectivity).

- *"By how much did Lehman Brothers' revenue increase in 2025 compared to 2024?"* (smooth temporal reasoning with unified result)
- *"How many years did it take Apple to release its first iPhone after its founding?"* (elegant temporal calculation)
- *"How many times larger is China's population compared to the U.S. population?"* (meaningful quantitative integration)

Bad Examples (Low Connectivity).

- *"What were Lehman Brothers' revenues in 2024 and 2025, respectively?"* (parallel enumeration, no synthesis)
- *"In what year was Apple founded, and in what year was the iPhone released?"* (choppy listing)
- *"What is the population of China, and when was the U.S. Constitution ratified?"* (superficial connection)

F.7.2 Criterion 2: Fluency

Purpose. Measures natural readability and conversational quality. Evaluates whether the question sounds like something a real person would naturally ask through two dimensions: (1) Naturalness—conversational authenticity without artificial template-like phrasing, and (2) Expression appropriateness—accessible vocabulary and straightforward syntax with creative rephrasing.

Good Examples (High Fluency).

- *"In what year did the female scientist who won the Nobel Prize twice receive the Nobel Prize in Chemistry?"*
- *"Which city is the headquarters of the world's largest e-commerce company founded by Jeff Bezos located in?"*
- *"How much did The Walt Disney Company pay to acquire the media group founded by Rupert Murdoch?"*

Bad Examples (Low Fluency).

- *"What is the temporal differential calculation when subtracting the institutional establishment year from the chronological marker of 1900?"* (unnecessarily complex)
- *"Identify the precise geographical placement of Amazon's principal global operational headquarters."* (artificial phrasing)
- *"State the capital monetary figure associated with Disney's large-scale corporate acquisition."* (overly formal)

F.7.3 Criterion 3: Essentiality

Purpose. Measures focus on core information while eliminating unnecessary elements. Evaluates modifier restraint—absence of excessive adjectives, adverbs, and decorative expressions that LLMs often generate.

Good Examples (High Essentiality).

- *"When did Marie Curie work at the University of Paris?"* (clean and factual)
- *"How much did Disney pay to acquire 21st Century Fox?"* (no unnecessary modifiers)
- *"What are the four bases that compose DNA?"* (direct and concise)

Bad Examples (Low Essentiality).

- *"What is the exact chronological period during which the highly distinguished Nobel laureate Marie Curie conducted her groundbreaking research at the prestigious University of Paris?"* (excessive modifiers)
- *"What was the exact amount of the historically significant deal when Disney announced its acquisition of the giant global media company 21st Century Fox?"* (decorative language)
- *"What are the four crucial fundamental elements that make up DNA, the intricate and sophisticated blueprint of genetics?"* (redundant descriptors)

F.7.4 Criterion 4: Validity

Purpose. Measures substantive worth and rational motivation. Evaluates through two dimensions: (1) Information worth—genuine insight beyond trivial fact-checking with educational/intellectual value, and (2) Query motivation—clear rational reason arising from natural curiosity or legitimate inquiry needs.

Good Examples (High Validity).

- *"What is considered Marie Curie's greatest achievement?"* (meaningful historical insight)
- *"Which scientists discovered the structure of DNA in 1953?"* (educational value)
- *"What renewable energy source currently has the highest efficiency rate?"* (practical knowledge)

Bad Examples (Low Validity).

- *"What is the exact number of letters in the name of the university where Marie Curie worked?"* (trivial numerical fact)
- *"What is the title of this document?"* (no substantive insight)
- *"On which page can the comprehensive report be found?"* (meta information, no value)

G Responsible NLP Checklist

G.1 For every submission

A1. Did you describe the limitations of your work?

Yes. Section: Limitations (after Conclusion).

A2. Did you discuss any potential risks of your work?

Yes. Section: Ethics Statement.

A3. Do the abstract and introduction summarize the paper's main claims?

Yes. Sections: Abstract and 1 (Introduction).

A4. Have you used AI writing assistants when working on this paper?

Yes. AI tools were used for grammar correction, proofreading, and English composition. However, all intellectual contributions, experimental design, and analysis were entirely conducted by humans.

G.2 Did you use or create scientific artifacts?

B1. Did you cite the creators of artifacts you used?

Yes. Model sizes are stated in all tables (e.g., Qwen3-0.6B, 4B, 8B, E5-Large-0.56B, BGE-M3-0.56B, Jina-v4-3.75B, E5-Mistral-7B). The total computational cost is reported in Appendix A.3 ($660.10). Since we did not perform any training, no additional computation cost was incurred; the inference infrastructure is described in Appendix A.2, and the associated cost is reported in Appendix A.5.

B2. Did you discuss the license or terms for use and release of your artifacts?

Yes. All source documents are publicly available: SEC 10-K filings (public disclosures), USPTO patents (public domain), U.S. Code (public law), Wikipedia (CC BY-SA). Sections 4 and Appendix A.1.

B3. Did you discuss if your use of existing artifacts is consistent with their intended use?

Yes. All artifacts used for their intended research purposes.

B4. Did you discuss the steps taken to check whether the data contains information that names or uniquely identifies individuals or offensive content?

Yes. Section 3.1 describes human evaluation validating LLM filtering. All source data is publicly available non-sensitive corporate/legal documents.

B5. Did you provide documentation of the artifacts?

Yes. Section 4 describes domain characteristics. Appendix A provides detailed corpus statistics, implementation details, and construction methodology.

B6. Did you report relevant statistics for the data you used/created?

Yes. Table 1 (main text) and Table 4 (Appendix A.2) report corpus statistics including chunk counts, redundancy metrics, and sample distribution across hop depths.

G.3 Did you run computational experiments?

C1. Did you report the number of parameters, computational budget, and computing infrastructure?

Yes. Model sizes stated in all tables (e.g., Qwen3-0.6B, 4B, 8B, E5-Large-0.56B, BGE-M3-0.56B, Jina-v4-3.75B, E5-Mistral-7B). Computational cost in Appendix A.3 ($660.10 total).

C2. Did you discuss how you selected hyperparameter values?

Yes. Appendix A.2 specifies all hyperparameters: chunking (512 tokens), embedding threshold ($\tau = 0.5$), top-k (k=3), multi-hop depth ($M \in \{1, 2, 3, 4\}$), candidate generation (N=100, S=10).

C3. Did you report descriptive statistics about your results?

Yes. Appendix E reports 5-run averages with mean, standard deviation, and CV% for SepRRF ablation.

C4. Did you provide access to code and data?

All data and code will be released publicly upon acceptance.

G.4 Did you use human annotators?

ß **D1. Did you report the full text of instructions given to participants?**

Partially. Section 3.1: "Each sample labeled by a single evaluator following identical criteria as the LLM"

D2. Did you report information about how you recruited and paid participants?

Yes. Section 3.1: "8 internal evaluators" who are "internal employees as part of regular duties."

D3. Did you discuss whether and how consent was obtained?

N/A. All data sources are publicly available documents. No personal data collected.

D4. Was the data collection protocol approved by an ethics review board?

No IRB approval required. Evaluation involved internal employees reviewing publicly available documents with no human subjects research.

D5. Did you report basic demographic and geographic characteristics of annotators?

No. Section 3.1 mentions 8 internal evaluators with domain expertise, though detailed demographic breakdown beyond professional background is not provided in the main text.