VERGE: Verification-Enhanced Generation of Multi-Hop Evaluation Datasets for Task-Specific RAG

Kyosuke Morita
Lloyds Banking Group
Kyosuke.Morita@lloydsbanking.com

Abstract

Evaluating retrieval-augmented generation (RAG) on custom knowledge bases requires task-specific benchmarks of multi-chunk evidence synthesis, yet most existing generation pipelines apply quality control only after the full question set has been solved. We present VERGE, which moves quality control inside the generation loop: a verification agent targets, per item and without an external solver pass, an operational multi-chunk criterion and a QA-integrity criterion, refining items that fail and discarding those it cannot repair. From 11,052 candidates across six domains, VERGE yields 2,854 verified multiple-choice items targeting aggregation- and comparison-style synthesis over topically coherent evidence. We contribute a method-agnostic validation protocol — exhaustive strict-subset ablation, chunk-removal ablation, format-unconfounded open-ended re-evaluation, and inter-chunk-similarity stratification — and use it to localise where the multi-chunk dependence holds. On a stratified sample (n = 250), 69.8% of items not answerable closed-book cannot be answered from any strict subset of their supporting chunks (strict-necessity violation rate 15.2%). A human annotation study finds the verifier reliable (Cohen's κ = 0.903) and VERGE items higher in QA integrity and distractor quality than the closest prior method ($p < 0.001$).

1 Introduction

Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2020) have become central to knowledge-intensive applications by combining parametric knowledge from large language models (LLMs) with non-parametric knowledge from external retrieval systems. Despite widespread industrial adoption, methodological challenges in evaluating RAG systems' multi-hop reasoning capabilities on custom enterprise knowledge bases persist.

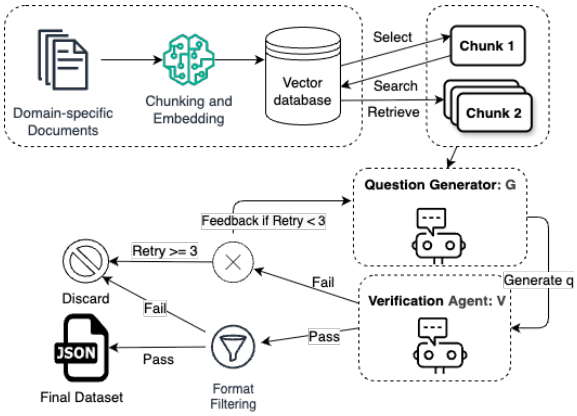


Figure 1: VERGE dataset generation process.

We introduce VERGE (VERification-enhanced GEneration),[1] a pipeline for generating task-specific evaluation datasets for aggregation- and comparison-style multi-hop QA, a subclass of multi-hop QA characterised in Section 3.1, for RAG systems. Its central design principle is integrating verification directly into the generation loop: a verification agent targets an operational multi-chunk criterion — no chunk subset should match full-context performance — and a QA-integrity criterion at generation time, and iteratively refines questions that fail those criteria, rather than applying filtering post-hoc (Guinet et al., 2024). We demonstrate empirically that verification accounts for the per-item quality gain while refinement contributes yield and coverage of the structurally hardest stratum; the quality gain holds across question formats (MCQ and open-ended) and across solver models, with verifier decisions exhibiting high agreement across five independent judges.

In summary, this paper makes the following contributions:

[1] Code, prompts and evaluation scripts: https://github.com/kyosek/VERGE

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3036–3062
November 6-10, 2026 ©2026 Association for Computational Linguistics

- A solver-free, per-item verification procedure for multi-hop RAG benchmark construction that operationalises a multi-chunk criterion and QA integrity at generation time.
- A method-agnostic validation protocol (exhaustive strict-subset ablation, chunk-removal ablation, format-unconfounded open-ended re-evaluation, inter-chunk-similarity stratification and a corpus-level shortcut audit) that diagnoses where multi-chunk dependence holds.
- Empirical evidence that verifier-based curation produces a more retrieval-discriminative instrument than unverified generation and than the closest prior method, validated against human judgement.
- A component analysis that separates the two mechanisms: a verification ablation attributing the entire per-item quality gain to verification, and a refined-vs-first-pass comparison showing that refinement contributes yield recovery — rescuing 11.4% of otherwise-discarded items without sacrificing quality.

2 Related Literature

A range of benchmark datasets have been proposed to evaluate multi-hop reasoning in question answering. Early datasets established foundational frameworks for evaluating multi-hop reasoning over multiple documents (Yang et al., 2018; Khashabi et al., 2018; Ho et al., 2020). Subsequent work extended this paradigm to the multiple-choice setting (Pang et al., 2022) and introduced more controlled compositional structures (Trivedi et al., 2022). Tang and Yang (2024) proposed MultiHop-RAG, a benchmark built from news articles for multi-hop queries in RAG contexts; like the datasets above, it is a fixed benchmark over a predetermined document collection rather than a method for generating task-specific evaluation instruments, which limits its use for RAG evaluation on specialised knowledge bases.

Among generation methods, RAGAs (Es et al., 2024) synthesises open-ended QA pairs from a document corpus with LLMs but applies no verification during generation, leaving question quality unchecked. Recent work explores compositional structure and difficulty calibration, but the majority of them rely on post-hoc filtering (Zhu et al., 2025; Shen et al., 2025; Lee et al., 2025b; Gong et al., 2025; Lee et al., 2025a; Zeng et al., 2025). Closest to our work, Guinet et al. (2024) generate multiple-choice exams for task-specific RAG assessment and optimise them post hoc. RAG systems first solve the generated questions, then Item Response Theory (IRT) is applied to evaluate exam quality and iteratively optimise the question set, so verification is postponed until after generation. In contrast, VERGE targets distractor-rich multiple-choice question generation for multi-hop RAG evaluation, integrating verification into the generation process. VERGE adapts established components (LLM-based generation, LLM-as-judge verification and the item-writing guidelines of Haladyna et al. (2002)); what is new is solver-free per-item acceptance at generation time and diagnosis-conditioned refinement from the verifier's stated cause of failure. Prior work has explored refinement of RAG components at response time (Shao et al., 2023; Asai et al., 2024; Zhou et al., 2024; Shi et al., 2024); we apply self-reflection to evaluation-dataset generation, and compare primarily against Guinet et al. (2024) as the closest prior work in multiple-choice RAG assessment.

3 VERGE framework

3.1 Problem Formulation

Our evaluation framework formulates the multi-hop RAG assessment problem as follows. Let $D = \{d_1, d_2, \ldots, d_n\}$ represent a corpus of documents where each document d_i contains potentially relevant information. We chunk D into semantic chunks $C = \{c_1, c_2, \ldots, c_m\}$, where each chunk c_j is a semantically coherent chunk. A multi-hop question q is a tuple $q = \langle question, A, a^*, C_q \rangle$, where $C_q \subseteq C$ with $2 \leq |C_q| \leq 5$ is the supporting set, a^* is the reference answer, and A is the answer space. By this definition, we do not target bridge reasoning, in which a semantically distant chunk must be located via an intermediate entity but rather q requires that the answer be derivable only by synthesising information distributed across multiple chunks of C_q. In practice, C_q is the top-k chunks most semantically similar to a randomly seeded chunk, which yields evidence sets that are distributed yet topically coherent. Mean pairwise cosine similarity within C_q is 0.80, against 0.50 for random within-domain pairs (about four standard deviations apart), so the clustering is construction-driven rather than an artefact of within-domain homogeneity. The construction therefore empha-

sises aggregation- and comparison-style synthesis over semantically proximate evidence and undersamples bridge-style chains; VERGE targets this subclass of multi-hop QA. For multiple-choice instantiation, $A = \{a_1, a_2, a_3, a_4\}$ with $a^* \in A$; for open-ended instantiation, A is unconstrained free text and a^* is a reference answer.

3.2 Question Generator

Given C_q and question generation instruction prompt, the generator G generates QA pair candidates in q. The answer candidates contain three distractors and one correct answer a^*. Whereas Guinet et al. (2024) require a separate process to obtain effective distractors, we integrate distractor engineering directly within the generation prompt. Following Haladyna et al. (2002), our prompt instructs the generator to produce three complementary distractor types targeting partial information reliance, contextual misalignment, and inter-chunk dependency failures (the prompt is in Appendix D.1).

3.3 Verification Agent

As illustrated in Figure 1, VERGE framework integrates the verification agent directly into the question generation process, distinguishing it from existing approaches that separate these stages. Our methodology employs a verification agent V that judges whether generated questions meet multi-hop requirements. During the question generation process, the verification agent verifies the question following two criteria: (1) **Complete Multi-Hop**: $V_c(q, C_q) \rightarrow \{0, 1\}$ judges whether deriving the correct answer a^* requires synthesising information distributed across multiple chunks in C_q, such that the question cannot be reliably answered without the full supporting context. The verifier does not provide a formal logical guarantee that no strict subset of evidence suffices; instead, it targets an operational criterion reflected by performance degradation under partial context. (2) **QA Integrity**: $V_v(q, C_q) \rightarrow \{0, 1\}$ judges whether a^* can be derived from information in C_q. When a question fails any verification criteria, the agent provides a refinement feedback designed to help the question meet the verification criteria. This process is repeated up to two refinement iterations before either meeting the verification criteria or discarding it, enabling the question generation process to iteratively refine itself. Additionally, we examine the question format to filter wrong format questions. If q does not meet the aforementioned format, q will be discarded.

Methodology	Timing	Refine.	Criteria	Feedback
Es et al. (2024)	None	None	None	None
Guinet et al. (2024)	Post	None	IRT model	Item stat. (pass/fail)
VERGE (Ours)	During	≤ 2 iter	Compl. MH and QA integ.	Per-item diagnosis

Table 1: Verification methodological differences between VERGE and existing task-specific evaluation dataset generation approaches. *Timing*: when verification occurs (Post = after generation, During = concurrent with generation); *Refine.*: number of iterative refinement cycles (iter = iterations); *Criteria*: evaluation metrics used; *Feedback*: signal returned per item .

Domain	# Question	# Chunk	Size	Voc. Count
Covid	422	270	332	5,522
GovReport	656	5,152	326	24,963
HotpotQA	328	5,433	347	74,440
MultiFieldQA	616	1,640	344	31,973
2Wiki	488	2,884	347	49,145
SEC Filings	344	36,834	337	52,968
Total	**2,854**	**52,213**	**338**	**140,648**

Table 2: Task domain breakdown of the benchmark dataset. *# Question*: number of questions; *# Chunk*: number of chunks in the corpus; *Size*: average chunk size (words); *Voc. Count*: unique vocabulary count.

Table 1 summarises the key methodological differences between VERGE and existing approaches for dataset generation.

Verifier choice and bias mitigation. We use Gemma2-9B (Riviere et al., 2024) as the primary verifier. Because LLM-as-judge evaluation can exhibit self-favouring bias (Zheng et al., 2023; Panickssery et al., 2024; Wang et al., 2024), generators span three model families (Llama 3.2-3B (Meta, 2024), Ministral-8B (Mistral AI, 2024), Gemma2-9B), so the accepted set is not tied to a single generation distribution, and verifier decisions are validated against three human annotators and four additional judges spanning open and frontier families.

3.4 VERGE Dataset

Our dataset consists of 2,854 multiple-choice multi-chunk questions across six different task domains: four datasets from LongBench datasets (Bai et al., 2024; Huang et al., 2021; Yang et al., 2018; Ho et al., 2020) and additionally incorporate UK Covid-19 Inquiry document (UK Covid-19 Inquiry, 2026) and SEC Edgar filings (Loukas et al., 2021).

Table 2 demonstrates the dataset statistics on different task domains. Our iterative verification framework yielded an acceptance rate of 26%, resulting in 2,854 verified questions from 11,052 candidates, with 326 questions successfully refined.

The 2,854 MCQ items constitute the primary VERGE dataset; two open-ended variants derive from it and are described together with their evaluation protocol in Section 4.1.2.

3.5 Dataset Quality Evaluation Criteria

To evaluate dataset quality, we conduct a human annotation study and instrument-level discriminativeness evaluation. Firstly, a human annotation study, performed by three independent ML practitioners blind to question source, serves four purposes across our experiments: (1) to validate the VERGE verifier's assessments against human judgement, (2) to empirically compare the VERGE dataset against Guinet et al. (2024) on QA integrity and distractor quality, (3) to characterise the quality difference between refined and non-refined questions (Section 5.2), and (4) to validate the GPT-4o judge used in the open-ended evaluation against human labels (Section 4.1.2).

Each annotator evaluates an identical set of QA pairs, presented in randomised order to reduce order effects, on (a) Complete Multi-Hop, (b) QA Integrity, and (c) distractor quality (1–5 Likert scale). We report Cohen's κ (Cohen, 1960) for verifier-human agreement, Fleiss' κ (Fleiss, 1971) (binary) and Krippendorff's α (Krippendorff, 2004) (nominal) for inter-annotator agreement, and Chi-squared (Pearson, 1900) / Mann-Whitney U (Mann and Whitney, 1947) for cross-dataset comparison. More details of the annotation study are in Appendix B.

Secondly, we assess instrument-level discriminativeness: the degree to which a benchmark's questions actually reward retrieval rather than being answerable from a model's parametric knowledge. We quantify this with the Oracle-Closed gap, defined as the difference in solver accuracy between the Oracle setting (question, answer candidates, and ground-truth supporting chunks C_q) and the Closed-book setting (question and answer candidates only, no retrieved context): $\Delta_{\text{O-C}} = \text{Acc}_{\text{Oracle}} - \text{Acc}_{\text{Closed}}$. A well-constructed retrieval benchmark should be difficult to answer without context (low Closed-book accuracy) and reliably answerable when the supporting evidence is supplied (high Oracle accuracy). A large $\Delta_{\text{O-C}}$ therefore indicates that performance is driven by the retrieved evidence rather than by memorised parametric knowledge. We report $\Delta_{\text{O-C}}$ averaged across the four solver models, and use it both to compare VERGE against the prior dataset and to attribute quality gains to individual pipeline components. We acknowledge that defining quality through solver performance introduces a degree of circularity. Therefore, we treat $\Delta_{\text{O-C}}$ as complementary to the human evaluation above and rely on the strict-subset and chunk-removal protocols for the orthogonal claim that the dependence is multi-chunk.

3.6 RAG Experiment Configuration

We employ four open language models; Llama3.1-8B (Grattafiori et al., 2024), Ministral-8B, Gemma2-9B and Qwen3.5-9B (Qwen Team, 2026) to solve the VERGE dataset. The model selection is guided by two criteria: architectural diversity and computational efficiency. For retrieval components, we evaluate five configurations: **Closed-book** and **Oracle** serve as a lower and upper bound of the experiment, **Dense:** Utilising bge-large-en-v1.5 (Xiao et al., 2024) embedding model with vector similarity search, **Sparse:** Implementing BM25 algorithm (Robertson and Zaragoza, 2009), and **Hybrid with Reranker:** Ensemble of dense and sparse retriever and reranking using ms-marco-MiniLM-L-2-v2 (Reimers and Gurevych, 2019; Wang et al., 2020) cross-encoder model. We explore three chunk size configurations (512, 2,048, and 4,000 characters) with top-k = 5, and evaluate both standard and Chain-of-Thought (CoT) prompting (Wei et al., 2022). We adapt the CoT prompts from Zhuang et al. (2024). For a multiple-choice format, we employ point-wise accuracy metric to evaluate the performance of the RAG pipelines. For an open-ended format, our primary measure is an LLM judge (GPT-4o (OpenAI et al., 2024)) that scores each generated answer for correctness against the reference answer a^*; we validate the judge against human annotation. Token-level F1 against the reference serves as a conservative surface-overlap lower bound, since it cannot register whether a fluent answer reflects multi-hop synthesis and is insensitive to valid paraphrase; BERTScore (Zhang et al., 2020) was excluded because its IDF-weighted overlap is largely decorrelated from multi-hop correctness and compresses the Oracle-Closed separation we aim to measure.

	Passed		Failed		κ	
Verifier	**MH**	**QA**	**MH**	**QA**	**hum**	**Gem**
Human (maj.)	95	97	4	66	–	0.90
Gemma2-9B	100	100	2	62	0.90	–
Llama3.2-3B	99	99	4	99	0.86	0.92
Ministral-8B	87	100	0	78	0.80	0.83
GPT-5	96	94	12	47	0.80	0.80
Claude Sonnet4.5	85	81	8	72	0.76	0.66

Table 3: Verifier robustness. Passed/Failed columns: % of human-passed / human-failed items each verifier accepts, MH = Complete Multi-Hop and QA = QA Integrity. κ: agreement with human majority and with Gemma2-9B.

Method	QA Integrity	Distractor Quality
Guinet et al. (2024)	89.3%	2.526 ± 1.047
VERGE (Ours)	97.0%	3.219 ± 0.683
stat / p-value	$12.8 / < 0.001$	$64{,}045.5 / < 0.001$

Table 4: Comparison of dataset quality between VERGE and Guinet dataset. QA Integrity reflects the proportion of QA pairs passing binary human evaluation. Distractor Quality is rated on a 1-5 Likert scale and reported with standard deviation. Statistical significance is determined via a Chi-squared test (QA Integrity) and Mann-Whitney U test (Distractor Quality).

	Token-F1		LLM-judge		Δ (O-C)	
Dataset	Closed	Oracle	Closed	Oracle	F1	LLM
RAGAs	0.52	0.59	74.0	79.9	0.07	5.9
VERGE-OE	0.40	0.42	65.9	83.0	0.02	17.1

Table 5: $\Delta_{\text{O-C}}$ on VERGE-OE (n = 1,290) and RAGAs (n = 1,075), averaged across four solver models. Token-F1: token-level F1 against the reference; LLM-judge: accuracy in % as rated by GPT-4o.

4 Experiment Results

4.1 Dataset Evaluation

4.1.1 Verifier Agreement

Table 3 presents human validation of the VERGE verification criteria: Complete Multi-Hop and QA Integrity. 95.0% of these questions met the Complete Multi-Hop criterion, 97.0% satisfied the QA Integrity criterion, and 93.3% passed both criteria. In contrast, among the questions that were rejected by VERGE verifier, only 4.0% satisfied the Complete Multi-Hop criterion and 65.7% satisfied QA Integrity. This large disparity shows that the VERGE verification agent effectively filters low-quality multi-hop questions.

Regarding agreement between human annotators and the verification agent, we observe Cohen's kappa of 0.903 for Complete Multi-Hop, 0.865 for QA Integrity, and 0.903 for overall pass/fail decisions. Fleiss' kappa scores were also high: 0.860 for Complete Multi-Hop, 0.813 for QA Integrity, and 0.864 for overall pass/fail decisions. Smaller open models (Llama3.2-3B, Ministral-8B) maintain high agreement with both human annotators and Gemma2-9B ($\kappa \geq 0.80$). Frontier models (GPT-5 (Singh et al., 2026) and Claude Sonnet 4.5 (Anthropic, 2025b)) also achieve substantial human agreement (κ = 0.80 and 0.76), yet accept more human-failed Complete Multi-Hop items (12% and 8%), i.e. they apply less conservative MH criteria; Claude Sonnet 4.5 also shows the lowest agreement with Gemma2-9B ($\kappa = 0.66$).

Against the human majority on the 200 annotated VERGE items, the verifier's false positives among accepted items are about 5% on Complete Multi-Hop and 3% on QA Integrity, and its false negatives among rejected items about 4% on Complete Multi-Hop. The dominant false-positive mode is reading topical distribution as informational distribution: when one chunk largely paraphrases another, a question drawing on both is answerable from either, the same mechanism that attenuates removal-induced degradation in the high-similarity stratum (Table 22).

4.1.2 Comparison to Prior Work

Table 4 presents a comparative analysis between VERGE dataset and Guinet dataset on two criteria: QA Integrity and Distractor Quality. The VERGE dataset achieves higher QA Integrity (97.0%) than the Guinet dataset (89.3%), with statistical significance confirmed by a Chi-squared test (χ^2 = 12.8, $p < 0.001$). Additionally, VERGE exhibits better overall distractor quality, averaging 3.219 ($\pm$ 0.683), compared to 2.526 ($\pm$ 1.047) for the Guinet dataset, as assessed via the Mann-Whitney U test (Mann and Whitney, 1947) (U = 64,045.5, $p < 0.001$). On these two quality metrics, the result is notable given that the Guinet dataset was generated using a more capable model than those employed in VERGE.

To examine whether the result depends on the multiple-choice format, we compare a natively generated open-ended VERGE variant (VERGE-OE; n = 1,290), produced by the same pipeline with the open-ended prompts of Appendix K, against RAGAs (Es et al., 2024) (n = 1,075) on the same six corpora, with the same four solvers under the

	Human annotation				Δ
Condition	**MH**	**QA**	**Both**	**Dist.**	**O−C**
No verif.	26.3	82.3	24.0	3.16 ± 1.01	9.2
Verif. only	94.3	97.7	93.6	3.24 ± 0.68	14.8
VERGE	95.0	97.0	93.3	3.22 ± 0.68	**15.1**

Table 6: Three-way ablation (n=300 items per condition for human annotation, n=273 for Verif. only as a subset of VERGE; three annotators). Human columns: % passing (MH, QA, Both) and mean Likert (Dist. = Distractor Quality). Δ: Oracle-Closed accuracy gap (pp), averaged across four RAG models, reported in Table 16. Inter-annotator agreement per condition is in Table 18.

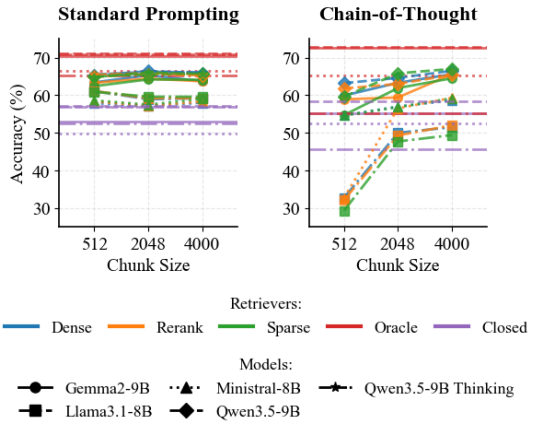


Figure 2: Accuracy in %. Results across four LLMs, three chunk sizes (512, 2,048, 4,000 characters), three retrievers, and two prompting strategies (Standard and Chain-of-Thought), plus closed-book and Oracle settings. For Qwen3.5-9B, Thinking denotes the thinking-enabled variant with standard prompting.

Closed-book and Oracle settings. Answers are scored by the GPT-4o judge, validated on 100 items against three annotators (Cohen's $\kappa = 0.89$, 95% CI [0.83, 0.94]; inter-annotator Fleiss' $\kappa = 0.86$), with token-F1 as a surface-overlap lower bound; a second open-ended evaluation, which strips the answer options from existing MCQ items, serves the format-agnostic ablation in Section 5.3. Table 5 demonstrates that VERGE-OE produces a substantially larger Oracle-Closed gap than RAGAs (17.1 pp vs. 5.9 pp), indicating that the retrieval dependence persists outside the multiple-choice format. Token-F1 gaps are small and similar across both datasets (0.02 vs. 0.07), consistent with the insensitivity of surface matching to free-form phrasing rather than with any absence of retrieval dependence.

4.1.3 Ablation: Effect of Verification

Table 6 reports the ablation of verification by comparing three conditions: no verification, verification only (first-pass acceptances), and the full VERGE dataset. Verification accounts for the entire per-item quality gain: Complete Multi-Hop adherence rises from 26.3% under no verification to 94.3% with verification, while QA integrity (82.3% to 97.7%), distractor quality (3.16 to 3.24), and the discriminative $\Delta_{\text{O-C}}$ (9.2 to 14.8 pp) reach the levels retained by the full pipeline.

The gain is concentrated in the multi-hop criterion: verification lifts Complete Multi-Hop adherence by roughly 68 pp but QA integrity by only 15 pp, because candidate generation already yields largely factually grounded QA pairs while routinely failing the multi-chunk requirement.

Every model's $\Delta_{\text{O-C}}$ widens under verification, and the effect is driven by improved Oracle answerability (mean 64.2% to 68.4%) rather than by suppressed closed-book accuracy (54.9% to 53.6%), indicating that verification makes questions reliably solvable from the supporting evidence rather than merely harder to answer from parametric knowledge (Table 16). The pattern holds for the two solvers that serve neither as generators nor as the verifier, Llama3.1-8B (8.1 to 12.6 pp) and Qwen3.5-9B (10.8 to 13.6 pp), so the widening is not attributable to self-favouring.

4.2 RAG Performance

Figure 2 reports RAG performance. First, almost all RAG configurations outperform the closed-book baseline and underperform the Oracle setting, which supports the utility of the VERGE dataset as a benchmark for multi-chunk evidence synthesis. Second, CoT prompting incurs a format-specific cost: its structured output fails to parse for a model-dependent fraction of items, and because these failures are not random with respect to difficulty, accuracies over parsed items only (Table 13) are optimistic relative to the full sample with failures counted as incorrect (Table 14). On the full-sample basis, standard prompting matches or exceeds CoT across chunk sizes, and CoT overtakes it among parsed items at the largest chunk size. Whether this reflects better use of longer chunks by more capable models or differential parse-failure selection is left to future work.

The controlled evidence for discriminativeness is the within-VERGE ablation of Table 6, in which $\Delta_{\text{O-C}}$ rises from 9.2 to 14.8 to 15.1 pp across

conditions that share the same corpus, generators, prompts and candidate pool. The cross-dataset comparison against the Guinet dataset (Table 15, Appendix H) is consistent with it: $\Delta_{\text{O-C}}$ is larger on VERGE than on the Guinet dataset by 6.1 (Ministral-8B) to 11.6 (Qwen3.5-9B) percentage points for every solver, and closed-book accuracy on Guinet exceeds that on VERGE by 21 to 29 pp, indicating that many Guinet questions are answerable from parametric knowledge or by eliminating obvious distractors. $\Delta_{\text{O-C}}$ is positive in all six domains, from 9.8 pp (2Wiki) to 21.3 pp (Covid).

5 Analysis and Discussion

5.1 Per-Criterion Filtering Analysis

The verification agent relies on two criteria to filter and refine the dataset: complete multi-hop and QA integrity. We observe asymmetry in the number of eliminated questions per verification criteria. The complete multi-hop criteria demonstrates higher filtering impact, eliminating 72.1% of candidates from the initial dataset. In contrast, the QA integrity criteria filters a comparatively modest 8.2% of questions, suggesting that generating a question with factual consistency is more achievable task than multi-hop reasoning requirements. The full results are reported in Table 9 in appendix I.

5.2 Refinement beyond Rejection Sampling

Mechanically, the refinement loop is rejection sampling with a diagnosis-conditioned proposal. A solver-based post-hoc filter returns pass or fail per item but the VERGE verifier additionally returns the cause of failure, most often a chunk in C_q that the question does not use, and the generator reproposes the item conditioned on that diagnosis (worked example in Figure 3, Appendix F). From the same 11,052 candidates, first-pass acceptance yields 2,528 items and refinement rescues a further 326 (11.4% of the final dataset; 260 at depth 1 and 66 at depth 2). In the human annotation study (100 unique items, three annotators), refined items exhibit higher hop complexity than first-pass items (mean required hops 3.49 vs. 2.98, with 50.0% vs. 31.3% of items requiring four or five hops; Mann–Whitney $U = 3892.5$, $p < 10^{-3}$), match them on Complete Multi-Hop validity (94.7% vs. 95.0%) and QA integrity (96.7% vs. 97.0%; $p = 0.85$ and $p = 0.81$), and show higher distractor quality (3.51 vs. 3.22; Mann–Whitney $p < 10^{-6}$; Table 17). The rescued items therefore populate the structurally hardest stratum rather than the borderline of the first-pass distribution, and the gain in distractor quality suggests the diagnosis-conditioned proposal also improves answer-option design. Finally, because acceptance is decided at construction time without a solver, the retained set is fixed independently of any solver panel, and the delta to Guinet et al. (2024), where the item set is selected through the responses of the systems under evaluation.

Outcome ($n = 250$)	**% items**
(a) Answerable closed-book	49.6
(b) Some strict subset of C_q suffices	15.2
(c) No strict subset of C_q suffices	35.2
(c) / ((b) + (c)): full C_q required	69.8

Table 7: Strict-subset ablation ($n = 250$; $3 \leq |C_q| \leq 5$; four solvers, $\geq 3/4$ rule). Percentages are of all 250 items; (b) and (c) partition the items not in (a).

5.3 Validating the Multi-Chunk Requirement

We test the verifier's Complete Multi-Hop judgement against solver behaviour in six ways.

Strict-subset ablation. On a sample of 250 items with $3 \leq |C_q| \leq 5$, stratified by domain and inter-chunk-similarity half, we solve each item under every strict subset of C_q (at most 30 per item) with the four solvers of Section 3.6, and call a condition answerable when at least three of the four are correct (chance $\approx 5\%$). Table 7 categorises the items: (a) 49.6% are answerable closed-book; (b) 15.2% are not answerable closed-book but are answerable from some strict subset, which we report as the strict-necessity violation rate; and (c) 35.2% are answerable only from the full C_q. Among items that require evidence at all, 69.8% require all of it (72.5% in the low-similarity stratum and 67.4% in the high-similarity stratum). Category (b) is an upper bound on the single-best-chunk answerability rate, since it counts subsets of any size, and category (a) reflects parametric answerability, which is orthogonal to evidence structure and is what $\Delta_{\text{O-C}}$ measures.

Chunk-removal ablation. At full scale, providing only the single best chunk lowers MCQ accuracy by 12.0% and removing one chunk from the full context by 5.5% on average relative to the Oracle setting (Table 19); the MCQ format partially compresses this gap, since models benefit from a 25% random baseline and can eliminate

implausible distractors from partial information. Re-evaluating 1,125 QA pairs open-ended, by removing the answer options and scoring token-level F1 and LLM-judge accuracy, gives single-chunk drops of 36.0% and 15.3% respectively, substantially larger than the MCQ figures: the multi-chunk dependence is not an artefact of the MCQ format but is partially masked by it in the primary experiments.

Cross-model stress test. A stress test using GPT-4o and GPT-4o-mini (OpenAI, 2024), a family used nowhere in the pipeline, evaluates 300 sampled QA pairs under closed-book, single-best-chunk, one-chunk-removed and Oracle conditions (Table 23). McNemar's tests (McNemar, 1947) show statistically significant degradation under all partial-context conditions for both models (all $p < 10^{-4}$), with 9–16 pp gaps between full-context and partial-context performance. The monotonic Closed $<$ Single Best $<$ One Less $<$ Oracle order holds for both models, consistent with the criterion capturing distributed evidence dependence rather than a verifier-specific artefact.

Similarity stratification. We stratify the eligible questions ($|C_q| \geq 3$; n = 1,679) at the median inter-chunk similarity (cosine = 0.83) into a low-similarity half (n = 839, range [0.36, 0.83]) and a high-similarity half (n = 840, range [0.83, 1.00]); both lie within VERGE's narrow high-similarity band, so this is a redundancy gradient rather than a bridge-reasoning test. In the low-similarity half, all four models degrade monotonically under chunk removal (single-best chunk: -3.3% to -10.1%; one chunk removed: -1.7% to -5.1%), the expected signature of distributed dependence (Table 22). In the high-similarity half, all four models also degrade, but Gemma2-9B and Llama3.1-8B are attenuated at single-best-chunk (-2.9% and -2.3%) while Ministral-8B and Qwen3.5-9B drop by -14.9% and -4.6%: redundancy reduces the marginal value of each additional chunk for models that can exploit partial context. The pooled degradation is therefore carried disproportionately by the low-similarity stratum; the multi-chunk requirement is strongest where inter-chunk redundancy is lowest.

Corpus-level shortcut audit. An item could still be answerable from a chunk outside C_q. We audit this directly: for each item we nominate the union of the top-10 non-C_q chunks returned by BM25, dense and hybrid-with-reranker retrieval, and apply the solver rule above to each nominated chunk on its own. 7.2% of items are solvable from some external chunk but not closed-book; on human inspection, 1.6% show consistent selection of the same distractor when an external chunk is supplied; and 95.2% of the items that require full context in Table 7 admit no corpus-level shortcut.

Lexical overlap. Finally, we identify for each item the chunk with the highest token-F1 overlap with the question and solve from that chunk alone. Accuracy is 52.1% overall and reaches 75.7% in the highest-overlap quartile (Tables 20 and 21), against a construction ceiling of 100% (every item passed QA Integrity), so lexical overlap explains at most part of the variance in single-chunk answerability and does not by itself resolve VERGE items. Together, these analyses show that the multi-chunk dependence survives exhaustive subset search, removal of the answer options and off-pipeline solvers, and is strongest where inter-chunk redundancy is lowest.

5.4 Generator Tradeoffs

Table 12 (Appendix G.4) reports wall time per refinement depth and monetary cost for all generators and for Claude 4.5 Sonnet and Haiku (Anthropic, 2025a): generation completes within 6–45 seconds per item depending on model size and depth, and API usage costs under $5 per 100 QA pairs for Sonnet and under $2 for Haiku. Wall time is end-to-end per item, since our logs do not separate generation, verification and refinement within an iteration; the evaluation side dominates total compute (36 retrieval configurations, two prompting strategies, closed-book and Oracle, over 2,854 items). To characterise the cost–yield tradeoff, we generated 200 QA pairs with a Claude pipeline (Sonnet 4.5 generator; Haiku 4.5 verifier) on the same domains and chunks: it achieves a higher acceptance rate (63% vs. 26%) and more first-pass acceptances (33% vs. 18%), reflecting higher initial generation quality, while both pipelines show a monotonic increase in average hop count with refinement depth, so the mechanism rescues structurally complex questions regardless of generator family and size.

6 Conclusion

We present VERGE, which moves quality control for multi-chunk RAG evaluation datasets inside the generation loop: an LLM-based verifier

targets an operational multi-chunk criterion and QA integrity at generation time, validated post hoc by chunk-removal and strict-subset ablation, enabling scalable curation with minimal human intervention. Verification yields a more retrieval-discriminative instrument than unverified generation, and diagnosis-conditioned refinement adds yield in the structurally hardest stratum without sacrificing quality.

Limitations

There are important inherent limitations to consider in this study. First, as both our method and experiments rely on LLMs, the quality of the dataset and experimental outcomes is directly influenced by the underlying model capabilities. Due to resource constraints, our experiments employ relatively small open models. This choice may lead to outcomes that differ from those obtained with larger state-of-the-art (SOTA) models, which typically exhibit substantially stronger generation and refinement abilities as previous sections collectively demonstrate. However, our human annotation study shows that the VERGE-generated dataset demonstrates higher-quality QA pairs compared to the Guinet dataset, which was generated using a larger model. Thus, this suggests that the VERGE method would remain effective and potentially generate even higher quality datasets when paired with larger SOTA models.

Second, because C_q is topically coherent by construction (Section 3.1), two consequences follow. The construction under-samples bridge-style chains, in which a semantically distant chunk must be located via an intermediate fact; we do not claim coverage of bridge reasoning, and results may not transfer to bridge-dominated benchmarks or to iterative, step-wise retrieval pipelines. Iterative retrievers are not disadvantaged by VERGE items, since a topically coherent C_q can be recovered in a single retrieval round, but VERGE does not measure the capability such retrievers exist for. Moreover, the strength of the multi-chunk dependence varies with inter-chunk redundancy: as our stratified ablation (Appendix J) shows, removal-induced degradation is present in both strata but is generally attenuated in the high-similarity stratum, with the degree of attenuation varying across models; the requirement is strongest where inter-chunk redundancy is lowest. We therefore recommend that multi-chunk ablations be reported stratified by inter-chunk similarity rather than pooled. We note separately that the verifier targets an operational notion of Complete Multi-Hop — no strict subset of chunks should match full-context performance, with systematic degradation under removal — rather than a formal guarantee that no subset suffices. The corpus-level shortcut audit is likewise conditional on the nominating retrievers, and no construction method, including human curation, certifies the absence of shortcuts (Min et al., 2019; Trivedi et al., 2020).

Third, while our experiments with CoT prompting demonstrated RAG performance improvements for specific model configurations when there is no parsing error, we did not explore these prompting techniques in the dataset generation framework. Given the observed effectiveness of such prompting strategies, integrating them into the dataset generation and verification process could potentially lead to higher-quality questions.

Fourth, evaluating dataset quality remains a fundamental methodological challenge. We define dataset quality in terms of discriminative power, yet this introduces potential circularity, as the dataset's value is inferred from the same model performance it is used to assess. This limitation parallels the IRT-based approach (Guinet et al., 2024), which optimises for informativeness post hoc via the IRT model rather than relying on independent quality metrics. The human annotation study remains the primary quality anchor; $\Delta_{\text{O-C}}$ is treated as complementary and is never used to rank RAG systems (Figure 2 ranks configurations within-solver).

Fifth, VERGE tests controlled evidence synthesis over a supplied or retrieved supporting set. It does not test behaviour under noisy retrieval, missing or conflicting evidence, stale information or ill-formed queries, so strong VERGE scores are not evidence of deployment reliability, particularly in high-stakes domains.

Acknowledgements

We thank all the reviewers for their constructive feedback and suggestions. Also, we thank the annotators for their time and effort. Finally, we thank Pedro Freire from Lloyds Banking Group for his valuable feedback on the paper.

References

Anthropic. 2024. Introducing Claude 3.5 Sonnet. Accessed: 2026-09-27.

Anthropic. 2025a. Introducing Claude Haiku 4.5. Accessed: 2026-09-27.

Anthropic. 2025b. Introducing Claude Sonnet 4.5. Accessed: 2026-09-27.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Albert Gong, Kamilundefined Stankevičiūtundefined, Chao Wan, Anmol Kabra, Raphael Thesmar, Johann Lee, Julius Klenke, Carla P. Gomes, and Kilian Q. Weinberger. 2025. Phantomwiki: on-demand datasets for reasoning and retrieval evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated evaluation of retrieval-augmented language models with task-specific exam generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16773–16801. PMLR.

Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Jeongsoo Lee, Daeyong Kwon, and Kyohoon Jin. 2025a. Grade: Generating multi-hop qa and fine-grained difficulty matrix for rag evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4405–4424.

Jeongsoo Lee, Daeyong Kwon, Kyohoon Jin, Junnyeong Jeong, Minwoo Sim, and Minwoo Kim. 2025b. Mhts: Multi-hop tree structure framework for generating difficulty-controllable qa datasets for rag evaluation. *arXiv preprint arXiv:2504.08756*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 13–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Accessed: 2026-09-27.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Mistral AI. 2024. Introducing Ministraux: The world's best edge models. Accessed: 2026-09-27.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. Accessed: 2026-09-27.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o system card. *arXiv*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Zhiyu Shen, Jiyuan Liu, Yunhe Pang, and Yanghui Rao. 2025. Hopweaver: Synthesizing authentic multi-hop questions across text corpora. *arXiv preprint arXiv:2505.15087*.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7339–7353, Bangkok, Thailand. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. OpenAI GPT-5 system card. *Preprint*, arXiv:2601.03267.

Yixuan Tang and Yi Yang. 2024. Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

UK Covid-19 Inquiry. 2026. Uk covid-19 inquiry module 4: Vaccines and therapeutics. UK Covid-19 Public Inquiry. Accessed: 2026-09-27.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

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, Brussels, Belgium. Association for Computational Linguistics.

Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. 2025. Rare: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *arXiv preprint arXiv:2506.00789*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1453–1463.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. RAGEval: Scenario specific RAG evaluation dataset generation framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. EfficientRAG: Efficient retriever for multi-hop question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3411, Miami, Florida, USA. Association for Computational Linguistics.

A Risk and Ethical Considerations

First, by operationalising "complete multi-hop reasoning" through a specific verifier-driven criterion, VERGE may bias evaluation toward particular reasoning patterns and inadvertently penalise alternative but valid strategies, contributing to benchmark-induced overfitting. Second, although we mitigate self-favouring effects by using multiple generator models, the LLM-based verifier may still encode architectural or prompt-specific biases that influence which questions are accepted or refined. To mitigate these risks, we position VERGE as a configurable evaluation instrument rather than a universal benchmark, encourage domain- and language-specific verifier adaptation, and release the generation, verification, refinement and evaluation code together with all prompts and criteria (footnote 1) to support transparency and auditing.

A.1 Licensing

The source corpora are derived from LongBench (MIT License; Bai et al. (2024)), UK Covid-19 Inquiry document (Open Government Licence v3.0 (OGL); UK Covid-19 Inquiry (2026)) and SEC Edgar filings (public domain, U.S. government documents) (Loukas et al., 2021). Also, all source datasets were used consistent with their intended

research purposes. The language models used - Gemma2-9B (Gemma Terms of Use), Llama 3.1/3.2 (Llama 3 Community License), Ministral-8B (Apache 2.0), Qwen3.5-9B (Apache License Version 2.0), GPT models (OpenAI Terms of Use) and Claude models (Anthropic Terms of Service) - permit research use.

B Human Annotation Study Details

The human annotation study was conducted to evaluate the quality of the QA pairs generated by VERGE compared to those from the Guinet et al. (2024) dataset. Annotators were machine learning practitioners who have MSc. degrees and were recruited through the authors' professional networks. None of the annotators were involved in the development of the method, experiment design, or paper writing. All annotations were performed independently following standardised written instructions, and annotators were compensated at a fixed hourly rate of £17 consistent with UK research assistant pay. Annotators were blind to the source of the questions and the identity of the generation method.

Annotation Sample Size per Study Purpose. The study serves the four purposes enumerated in Section 3.5: (1) validating the VERGE verifier's decisions against human judgement, (2) comparing the VERGE dataset against Guinet et al. (2024) on QA integrity and distractor quality, (3) characterising the quality difference between refined and non-refined questions, and (4) validating the GPT-4o judge used in the open-ended evaluation. For purposes (1) and (2), annotators evaluate 300 QA pairs: 100 from VERGE-passed, 100 from VERGE-failed, and 100 from the Guinet dataset. The 200 VERGE items (passed and failed) additionally serve as the human reference for verifier robustness validation, where we compare Gemma2-9B against four additional LLM verifiers: Llama3.2-3B, Ministral-8B, GPT-5, and Claude Sonnet 4.5. For purpose (3), annotators evaluate an additional 100 QA pairs sampled from refined and non-refined VERGE questions. For purpose (4), annotators label 100 open-ended responses to validate the GPT-4o judge, with agreement reported in Section 4.1.2.

B.1 Annotation Instruction

Task Overview

You will be shown a set of multiple-choice question-answer (QA) items generated for evaluating multi-hop reasoning in Retrieval-Augmented Generation (RAG) systems. Each QA item consists of:

- a question,
- four answer options (one correct answer and three distractors),
- a set of supporting document chunks.

Your task is to evaluate the quality of each QA item according to the criteria described below. You should base your judgments only on the information provided in the question and the accompanying document chunks. Please do not rely on external knowledge beyond what is explicitly stated in the provided context.

1. Verification Criteria

For each QA item, you will assess two binary criteria.

(a) Complete Multi-Hop Reasoning

Determine whether the question requires synthesising information from multiple document chunks in order to reliably identify the correct answer. A question satisfies Complete Multi-Hop if:

- The correct answer cannot be reliably determined using information from a single chunk alone.
- Multiple chunks must be jointly considered to resolve the question.

A question does not satisfy this criterion if:

- One chunk alone is reliably sufficient to answer the question correctly.
- The question can be answered using general knowledge without integrating the provided chunks.

Please answer Yes or No.

(b) Question-Answer (QA) Integrity

Determine whether the designated correct answer is fully supported by the provided document chunks. A question satisfies QA Integrity if:

- The correct answer is unambiguously derivable from the provided chunks.
- No factual inconsistencies or unsupported claims are present.

A question does not satisfy this criterion if:

- The correct answer contradicts the provided context.
- Key information required to justify the correct answer is missing.

Please answer Yes or No.

2. Distractor Quality Assessment

For each QA item, evaluate the quality of the three incorrect answer options (distractors).

Rate the distractors on a 5-point Likert scale, considering:

- Relevance: How closely the distractors relate to the question and context.
- Plausibility: Whether the distractors appear reasonable given partial or incorrect reasoning.
- Difficulty: Whether the distractors would challenge a model or reader who has not fully integrated all required information.

Scale definition:

- 1 - Very poor (irrelevant or obviously incorrect)
- 2 - Poor
- 3 - Good
- 4 - Very Good
- 5 - Excellent (highly plausible and challenging)

Please assign one the score per item.

3. Annotation Guidelines

- Work independently and do not discuss your evaluations with others.
- If unsure, select the option that best reflects your judgment.
- Please focus on accuracy and consistency.

All responses will be anonymised and used solely for research purposes.

C Use of AI assistants

We used AI-assisted tools to support aspects of writing refinement and routine code implementation. Specifically, these tools were employed to improve clarity, grammar, and to draft portions of straightforward code that can be reliably specified by explicit instructions. All AI-assisted outputs were carefully reviewed, edited, and validated by the authors. The authors retain full responsibility for the scientific content, experimental design, correctness of the implementation, and all claims made in this paper. The use of AI assistants did not affect authorship attribution.

D VERGE Prompts

D.1 Question Generation Prompt

Advanced Multi-Hop Question Generation Methodology Objective:
Generate an exceptionally challenging multiple-choice question (1 correct answer and 3 distractors) that demands sophisticated, interdependent reasoning across multiple document chunks.

Comprehensive Question Design Criteria:

1. **Multi-Hop Reasoning Architecture**
   - Construct questions that necessitate intricate information synthesis
   - Enforce mandatory cross-chunk logical dependencies
   - Eliminate single-chunk solution pathways
   - Create a complex reasoning landscape that requires:
     - Temporal reasoning
     - Contextual cross-referencing
     - Nuanced inference mechanisms
     - Sophisticated information integration
2. **Chunk Interdependency Strategies**
   - Mandate that NO single document chunk reliably provides a complete answer
   - Implement strategic information distribution across chunks
   - Create deliberate cognitive gaps requiring active reasoning
   - Establish intricate logical bridges between disparate information sources

3. **Distractor Engineering Principles**
   Three distractors must be meticulously crafted to exploit specific reasoning vulnerabilities:
   a) **Partial Information Distractor**
      - Leverages incomplete chunk understanding
      - Appears superficially plausible
      - Represents a common cognitive shortcut
   b) **Contextual Misalignment Distractor**
      - Based on seemingly logical but fundamentally flawed reasoning
      - Requires subtle comprehension of inter-chunk relationships
      - Exploits potential misinterpretation patterns
   c) **Critical Dependency Distractor**
      - Appears correct if crucial inter-chunk dependencies are overlooked
      - Demands highest-order reasoning to distinguish from correct answer
      - Represents a sophisticated reasoning challenge

4. **Reasoning Complexity Calibration**
   - Questions should require:
     – Minimum {num_chunks}-step reasoning process
     – Explicit integration of ALL provided chunks
     – Ability to distinguish nuanced information relationships
   - Difficulty Level: Advanced (Graduate/Professional Examination Standard)

5. **Linguistic Precision Requirements**
   - Use sophisticated, academically rigorous language
   - Avoid direct chunk referencing
   - Construct questions requiring active cognitive engagement
   - Ensure questions are solvable exclusively through provided documentation

Output Format:
Question: [Sophisticated multi-hop reasoning challenge]
A) [Option A]
B) [Option B]
C) [Option C]
D) [Option D]
Correct Answer: [Letter one of "A", "B", "C" or "D"]
* *Do not generate anything else other than the above output format*

Fundamental Constraints:

- Zero reliable single-chunk solution possibilities
- Mandatory full documentation utilization
- Rigorous reasoning complexity
- Linguistically sophisticated presentation

Evaluation Heuristic:
A high-quality multi-hop question successfully prevents:

- Surface-level information scanning
- Premature answer selection
- Incomplete reasoning processes

Domain: {task_domain}
Documentation: {documentation}

Generate an advanced multi-hop reasoning question

D.2 Verification Agent Prompt

Advanced Multi-Hop Question Verification Methodology Objective:
You are an advanced exam question verifier specializing in multi-hop reasoning and comprehensive document analysis.

Verification Objectives:

1. **Multi-Hop Reasoning Verification**
   - Determine if the question requires synthesizing information across multiple document chunks
   - Identify all necessary chunks for a comprehensive answer
   - Detect and prevent "shortcut reasoning" that bypasses critical information

2. **Question-Answer Integrity Assessment**

- Evaluate the coherence and correctness of the question and all options
- Verify that the correct answer is truly supported by the given documents
- Assess the depth and complexity of reasoning required

3. **Chunk Utilization Analysis**
    - Confirm whether all provided document chunks are essential
    - Identify any unused or potentially relevant chunks
    - Recommend strategies to incorporate underutilized information

Output Format:
Generate a comprehensive JSON response with the following mandatory entities:

- `required_chunks`: List of chunk indices essential for answering
- `reasoning`: Dictionary containing:
    - `unused_chunks`: List of chunks not utilized in the answer
    - `qa_integrity`: Boolean for Question-Answer Integrity Assessment
    - `improvement_suggestions`: Recommendations for exam enhancement
- `confidence`: 1-5 confidence scale in the assessment

JSON Schema:

```
{
   "required_chunks": [int],
   "reasoning": {
      "unused_chunks": [int],
      "qa_integrity": bool,
      "improvement_suggestions": str
   },
   "confidence": int
}
```

Specific Verification Criteria:

- **Multi-Hop Complexity:** Assess if answering requires integrating information from multiple sources
- **Information Completeness:** Evaluate whether all relevant document information is meaningfully incorporated
- **Reasoning Depth:** Determine the level of analytical thinking needed to derive the correct answer

Guidance for Analysis:

- Critically examine each document chunk's contribution
- Highlight potential ambiguities or gaps in the question-answer construct
- Provide constructive feedback for improving exam quality and multi-hop reasoning challenge

Input Variables:
Question: {question}
Options:
A) {option_a}
B) {option_b}
C) {option_c}
D) {option_d}
Correct Answer: {correct_answer}
Documents: {chunk_text}

Fundamental Constraints:

- Generate only JSON response following the specified schema
- Provide comprehensive analysis of multi-hop reasoning requirements
- Ensure rigorous evaluation of question-answer integrity
- Maintain focus on document chunk utilization assessment

Perform comprehensive multi-hop reasoning verification

D.3 Standard Prompt

Multiple Choice Exam Taking Methodology Objective:
You are an AI assistant taking a multiple choice exam. Your task is to analyze the provided materials and select the most appropriate answer.

Examination Process:

1. **Comprehensive Document Analysis**

- Read the question, provided choices and documents carefully
- Extract relevant information from all provided documentation
- Identify key concepts and relationships within the materials

2. **Choice Evaluation Framework**
   - Analyze each choice systematically against the documentation
   - Assess the logical consistency of each option
   - Identify the choice with strongest documentary support
3. **Answer Selection Protocol**
   - Select the most appropriate answer based on evidence
   - Ensure selection aligns with provided documentation
   - Apply rigorous analytical reasoning to final choice

Input Variables:
Question: {question.question}

Choices:
{formatted_choices}

Documents:
{question.documentation}

Response Format Requirements:

- You must respond with exactly one letter: A, B, C, or D
- Do not include any explanation, period, or additional text
- Just the letter of the correct answer

Examples of Valid Responses:

- A
- B
- C
- D

Fundamental Constraints:

- Generate only single letter response
- Base selection exclusively on provided documentation
- Maintain strict adherence to response format
- Ensure analytical rigor in choice evaluation

Your answer (one letter only):

D.4 Chain-of-Thought prompt

Step-by-Step Reasoning Methodology Objective:
As an assistant, your task is to answer the question after `<Question>` by using the given supporting documents. You should first think step by step about the question and give your thought, then answer the `<Question>`.

Response Format Requirements:

- Your answer should be after `<Answer>` in JSON format
- Use key "thought" and "answer" with string values
- The "answer" must be one of "A", "B", "C", or "D"
- Do not output anything else

Reasoning Process:

1. **Document Analysis**
   - Systematically examine the supporting documents
   - Extract relevant information for the given question
   - Identify key facts and relationships
2. **Step-by-Step Thinking**
   - Break down the question into component parts
   - Connect document information to question requirements
   - Apply logical reasoning to derive conclusions
3. **Answer Selection**
   - Evaluate each choice against your reasoning

- Select the most appropriate option
- Ensure selection is supported by documentary evidence

Example Responses:

Example 1:
<Question>: What is the name of this American musician, singer, actor, comedian, and songwriter, who worked with Modern Records and born in December 5, 1932?
<Answer>:
{{"thought":"Modern Record is a big R&B label with artists including Etta James, Joe Houston, Little Richard, Ike, Tina Turner and John Lee Hooker in the 1950s and 1960s. Little Richard is an American musician, signer actor and songwriter, born in December 5 1932. So the answer is Little Richard.","answer": "Little Richard"}}

Example 2:
<Question>: Between Chinua Achebe and Rachel Carson, who had more diverse jobs?
<Answer>:
{{"thought":"Chinua Achebe was a Nigerian novelist, poet, professor, and critic. Rachel Carson was an American marine biologist, author, and conservationist. Chinua Achebe has 4 jobs while Rachel Carson has 3 jobs. So the answer is Chinua Achebe.", "answer": "Chinua Achebe"}}

Example 3:
<Question>: Remember Me Ballin' is a CD single by Indo G that features an American rapper born in what year?
<Answer>:
{{"thought":"Remember Me Ballin' is the CD singer by Indo G that features Gangsta Boo, who is named Lola Mitchell, an American rapper born in 1979. So the answer is 1979.", "answer": "1979"}}

Input Variables:
Supporting Document:
{question.documentation}

Question:
<Question>: {question.question}
<Choices>: {formatted_choices}
<Answer>:

Fundamental Constraints:

- Generate only JSON response with specified format
- Provide comprehensive step-by-step reasoning in "thought" field
- Select answer exclusively from provided choices (A, B, C, or D)
- Base all reasoning on supporting documentation

Provide step-by-step reasoning and answer selection

E Question Examples

Below is one arbitrarily selected question per task domain.

E.1 Example GovReport Question

Question:
Considering the multifaceted approaches to addressing Puerto Rico's fiscal challenges outlined in the provided documentation, what is the most significant factor hindering the successful implementation of long-term economic solutions?

Choices:

A) The lack of a comprehensive, integrated information technology system for financial management.

B) The Oversight Board's reluctance to certify updated fiscal plans due to the impact of Hurricanes Irma and Maria.

C) The Puerto Rican government's dependence on federal assistance and its inability to generate sufficient revenue.

D) The absence of a clear consensus among federal agencies regarding the most effective forms of technical assistance.

Correct Answer: C

E.2 Example HotpotQA Question

Question:
Which metropolitan area, based on commuting patterns, encompasses the largest population within the broader "Gold Coast" region of Florida?

Choices:

A) Miami-Fort Lauderdale-Miami Beach Metropolitan Statistical Area

B) West Palm Beach-Boca Raton-Boynton Beach Metropolitan Statistical Area

C) Port Saint Lucie MSA

D) Sebastian-Vero Beach MSA

Correct Answer: A

E.3 Example MultiFieldQA Question

Question:
Given the Shadoks' optimization strategies and the BDFS algorithm's purpose, what is the most likely reason the Shadoks observed a 30 solution improvement near the end of the challenge after incorporating BDFS?

Choices:

A) BDFS effectively identified and eliminated redundant color classes, leading to a more efficient coloring scheme.

B) BDFS's ability to recolor vertices at larger distances from the conflict vertex allowed for a more nuanced exploration of the color space.

C) BDFS's conflict minimization phase directly addressed the limitations of the Shadoks' initial conflict set management strategy.

D) BDFS's use of a Gaussian random variable for the weight function (f) significantly improved the exploration of the solution space.

Correct Answer: B

E.4 Example SecFilings Question

Question:
Considering the evolution of U.S. assistance to Georgia, what was the primary shift in focus observed between the period immediately following the 2008 war and the years from FY2010 to FY2017?

Choices:

A) A transition from humanitarian aid to long-term economic development initiatives.

B) A shift from military capacity building to promoting democratic reforms and civil society.

C) A move from direct budgetary support to funding specific security and stabilization programs.

D) A change from supporting Georgia's participation in international peacekeeping missions to bolstering its border security.

Correct Answer: C

E.5 Example 2Wiki Question

Question:
What motivated Louisa Ulrika's opposition to the marriage between Gustav and Sophia Magdalena?

Choices:

A) She believed Sophia Magdalena was unsuitable due to her perceived lack of intelligence.

B) She desired a political alliance with Brandenburg-Schwedt through a marriage between Gustav and her niece.

C) She resented the Swedish public's preference for Sophia Magdalena over her own preferred candidate.

D) She felt the Swedish Queen exerted undue influence over the decision-making process.

Correct Answer: B

	Gemma2-9B	Llama3.2-3B	Ministral-8B
Generated	3,670	3,712	3,670
QAs	1,025	822	1,007
Success	27.9	22.1	27.4

Table 8: Benchmark dataset question survival rate over each LLM generators. Generated indicates the number of generated question-answer pairs before verification. QAs indicates the number of question-answer pairs that passed the verification process. Success indicates the percentage of question-answer pairs that passed the verification process.

Criteria	**# eliminated**	**% eliminated**
Complete Multi-Hop	7,967	72.1%
QA Integrity	902	8.2%
Both Criteria	7,885	71.3%
Format Filtering	313	2.8%

Table 9: Number of eliminated questions per verification criteria and final formatting filtering. Initially there are 11,052 candidate questions and survived 2,854 questions.

F Verification Process Survival Rate

Table 8 presents the survival rates of generated question-answer pairs across the three LLMs used for generation.

The number of surviving questions also varies across task domains: survival is lowest for HotpotQA, whose source documents require synthesising information across multiple documents, and highest for GovReport, where generated questions are more straightforward.

Table 9 provides a detailed breakdown of the elimination process during verification, showing the number and percentage of questions eliminated based on each criterion. The majority of questions were eliminated due to failing the complete multi-hop requirement, which underscores the challenge of generating questions that necessitate synthesizing information from multiple chunks. Additionally, a significant portion of questions failed the QA integrity criterion, indicating issues with solvability using only the provided chunks. The final formatting filtering also contributed to a small percentage of eliminations, ensuring that all surviving questions adhere to the required format for evaluation.

G RAG Implementation Details for Reproducibility

G.1 Model Sources

We list all the model versions we used in our experiments below in the table 10.

G.2 Model Parameters

For reproducibility purpose, we list language model parameters in the table 11.

G.3 Computational Resources

All the VERGE experiments were conducted on a MacBook Pro equipped with an Apple M3 Max processor and 48 GB of unified memory. GPU-accelerated computations were performed using the integrated Apple Silicon GPU. Guinet et al. (2024) dataset was generated by using claude-3-5-sonnet@20240620 (Anthropic, 2024) through Anthropic API on Vertex AI.

G.4 Cost and Generator Tradeoffs

Table 12 presents the average wall time per question-answer pair generation across different verifier depths and models.

H RAG results details

Table 13 reports the full results of the RAG experiment over successfully parsed items, with the number of parsed items in parentheses; because of parsing errors, each reported result has a different total.

Table 14 reports the CoT results over the full sample of 2,854 items with parse failures counted as incorrect.

Table 15 presents the RAG performance comparison between VERGE and Guinet dataset across five models (Gemma2-9B, Llama3.1-8B, Ministral-8B, Qwen3.5-9B, Claude Sonnet 4.5) and four retrieval configurations (Closed-book, Sparse, Dense, Oracle). Values represent accuracy in %.

Table 16 presents the RAG performance on the ablation study on "no verification", "verification only" and "VERGE" settings across four models (Gemma2-9B, Llama3.1-8B, Ministral-8B, Qwen3.5-9B) and two retrieval configurations (Closed-book, Oracle). Values represent accuracy in %.

I Ablation Study on Verifier

Table 17 presents the evaluation results of refined and non-refined questions comparison in the verifi-

Before Refinement:
Q: Under what specific circumstances would Broadjam be obligated to remove user-submitted Material from public view?
Answer Candidates:
A) When a user requests removal of their own Material
B) When Broadjam receives a valid claim of copyright infringement
C) When a user reports that Material violates Broadjam's Terms of Service
D) When Broadjam determines that Material is harmful or offensive
Feedback: While Chunk 0 provides the necessary information, Chunk 1 could be incorporated for a more comprehensive understanding
Required Chunks: Chunk 0

After Refinement:
Q: A user believes their copyrighted material was posted on Broadjam without permission. What steps should they take, according to Broadjam's policies?
Answer Candidates:
A) Immediately contact Broadjam's Copyright Agent and request removal.
B) File a formal copyright claim with the US Copyright Office and notify Broadjam.
C) Contact the user who posted the infringing material directly.
D) Register with the US Copyright Office and then notify Broadjam.
Feedback: The question and answer are well-aligned with the provided documents.
Required Chunks: Chunk 0 and 1

Figure 3: Example of question refinement process showing improvement in satisfying the verification criteria.

Model	Release	Model Path and File Name
Gemma2-9B	2024-07	bartowski/gemma-2-9b-it-GGUF (gemma-2-9b-it-Q4_K_M.gguf)
Llama3.1-8B	2024-07	bartowski/Meta-Llama-3-8B-Instruct-GGUF (Meta-Llama-3-8B-Instruct-Q4_K_M.gguf)
Llama3.2-3B	2024-09	hugging-quants/Llama-3.2-3B-Instruct-Q8_0-GGUF (*q8_0.gguf)
Ministral-8B	2024-10	bartowski/Ministral-8B-Instruct-2410-GGUF (Ministral-8B-Instruct-2410-Q4_K_M.gguf)
GPT-4o	2024-05	gpt-4o
GPT-4o-mini	2024-07	gpt-4o-mini
GPT-5	2025-08	gpt-5
Claude Sonnet 3.5	2024-06	claude-3-5-sonnet@20240620
Claude Sonnet 4.5	2025-09	claude-sonnet-4-5
Claude Haiku 4.5	2025-09	claude-haiku-4-5
Qwen3.5-9B	2026-03	unsloth/Qwen3.5-9B-GGUF (Qwen3.5-9B-Q4_K_M.gguf)

Table 10: List of LLMs used in our experiments

Model	Temperature	top-p
Gemma2-9B	0.0	0.9
Llama3.1-8B	0.0	0.9
Llama3.2-3B	0.0	0.9
Ministral-8B	0.0	0.9
GPT-4o	0.0	0.9
GPT-4o-mini	0.0	0.9
GPT-5	0.0	0.9
Claude Sonnet 3.5	0.0	0.9
Claude Sonnet 4.5	0.0	0.9
Claude Haiku 4.5	0.0	0.9
Qwen3.5-9B	0.0	0.9

Table 11: Hyperparameters of the models.

Model	Iter 0	Iter 1	Iter 2	Cost
Llama3.2-3B	6s	15s	27s	-
Gemma2-9B	14s	27s	41s	-
Ministral-8B	13s	28s	41s	-
Claude 4.5 Sonnet	15s	29s	45s	4.45
Claude 4.5 Haiku	12s	25s	40s	1.87

Table 12: Average wall time per QA pair generation across different verifier depths and models. Cost is calculated based on the API pricing as of Nov 2025 for Claude 4.5 per 100 QAs in USD.

cation process.

Figure 3 illustrates the mechanism on a single-chunk question that becomes a genuine multi-hop item once an unused chunk is incorporated. A second case from the SEC domain shows the same pattern at higher complexity: a question initially answerable directly from two filing paragraphs was refined to require cross-referencing three, raising $|C_q|$ from two to three.

Table 18 reports inter-annotator agreement for the three conditions of the verification ablation (Table 6).

J Details of Complete Multi-Chunk Validation

Table 19 presents the complete ablation study results on multi-hop questions.

Table 20 and 21 present the results of lexical

		Gemma2-9B		Llama3.1-8B		Ministral-8B		Qwen3.5-9B		
Retriever	**Chunk**	**Standard**	**CoT**	**Standard**	**CoT**	**Standard**	**CoT**	**Standard**	**CoT**	**Thinking**
Closed	-	53.1 (2,854)	55.1 (2,643)	52.6 (2,854)	45.6 (2,456)	49.9 (2,854)	52.5 (2,406)	57.1 (2,854)	58.3 (2,722)	56.9 (2,854)
Oracle	4000	70.3 (2,854)	72.6 (2,604)	65.2 (2,854)	55.1 (2,376)	66.5 (2,854)	65.3 (2,421)	71.1 (2,854)	72.8 (2,688)	71.0 (2,854)
Dense	512	63.4 (2,854)	59.9 (2,271)	**61.1** (2,854)	32.7 (1,847)	57.9 (2,854)	54.6 (2,320)	65.2 (2,854)	63.2 (2,667)	65.5 (2,854)
	2048	**65.3** (2,854)	63.3 (2,354)	58.9 (2,854)	50.0 (2,217)	57.1 (2,854)	57.0 (2,418)	**66.4** (2,854)	64.6 (2,699)	**66.2** (2,854)
	4000	63.7 (2,854)	**65.6** (2,445)	59.2 (2,854)	51.5 (2,231)	58.0 (2,854)	58.6 (2,311)	65.8 (2,854)	66.5 (2,702)	66.1 (2,854)
Rerank	512	63.3 (2,854)	58.9 (2,355)	**61.1** (2,854)	32.1 (1,920)	58.3 (2,854)	32.7 (2,418)	64.9 (2,854)	61.8 (2,628)	65.4 (2,854)
	2048	64.4 (2,854)	59.4 (2,464)	59.2 (2,854)	49.3 (2,052)	57.2 (2,854)	56.5 (2,482)	65.6 (2,854)	63.0 (2,693)	65.9 (2,854)
	4000	63.7 (2,854)	**65.6** (2,407)	59.2 (2,854)	**52.0** (2,049)	58.3 (2,854)	**59.3** (2,321)	65.3 (2,854)	65.4 (2,602)	65.5 (2,854)
Sparse	512	62.4 (2,854)	54.7 (2,332)	60.9 (2,854)	29.3 (1,421)	58.6 (2,854)	54.6 (2,393)	65.1 (2,854)	59.6 (2,589)	64.8 (2,854)
	2048	64.2 (2,854)	61.9 (2,635)	59.6 (2,854)	47.7 (1,992)	57.5 (2,854)	56.8 (2,401)	65.5 (2,854)	65.8 (2,611)	66.0 (2,854)
	4000	64.1 (2,854)	64.5 (2,573)	59.6 (2,854)	49.4 (1,922)	**59.0** (2,854)	59.2 (2,429)	65.7 (2,854)	**67.0** (2,712)	65.7 (2,854)

Table 13: Accuracy in % with sample sizes in parentheses. Results across four LLMs, three chunk sizes (512, 2048, 4000), three retrievers, and two prompting strategies (Standard, CoT: Chain-of-Thought), plus closed-book and Oracle settings. For Qwen3.5-9B, Thinking denotes the thinking-enabled variant with standard prompting. The best configurations per model and prompt are indicated by bold.

Retriever	**Chunk**	**Gemma2-9B**	**Llama3.1-8B**	**Ministral-8B**	**Qwen3.5-9B**
Closed	-	51.0	39.2	44.3	55.6
Oracle	4000	66.2	45.9	55.4	68.6
Dense	512	47.7	21.2	44.4	59.1
	2048	52.2	38.8	48.3	61.1
	4000	56.2	40.3	47.5	63.0
Rerank	512	48.6	21.6	27.7	56.9
	2048	51.3	35.4	49.1	59.4
	4000	55.3	37.3	48.2	59.6
Sparse	512	44.7	14.6	45.8	54.1
	2048	57.2	33.3	47.8	60.2
	4000	58.1	33.3	50.4	63.7

Table 14: Accuracy of RAG with CoT prompting with the whole sample in %. Results across four LLMs, three chunk sizes (512, 2048, 4000), three retrievers, plus closed-book and Oracle settings. The sample size is 2,854 for all the experiments.

Model	Gemma2-9B		Llama3.1-8B		Ministral-8B		Qwen3.5-9B		Sonnet 4.5	
Dataset	Guinet	VERGE	Guinet	VERGE	Guinet	VERGE	Guinet	VERGE	Guinet	VERGE
Closed	80.4	52.2	73.6	52.4	74.9	49.3	83.2	54.0	84.1	58.3
Sparse	79.9	63.1	71.2	59.2	74.2	59.0	82.7	62.3	87.5	69.8
Dense	84.1	63.6	75.1	58.8	77.8	57.9	85.0	62.4	91.0	70.4
Oracle	89.4	69.7	79.5	64.8	85.5	66.0	88.0	70.4	94.0	77.9
$\Delta_{\text{O-C}}$	9.0	17.5	5.9	12.4	10.6	16.7	4.8	16.4	9.9	19.6

Table 15: RAG performance comparison between VERGE and Guinet dataset across five models (Gemma2-9B, Llama3.1-8B, Ministral-8B, Qwen3.5-9B, Claude Sonnet 4.5) and four retrieval configurations (Closed-book, Sparse, Dense, Oracle). Values represent accuracy in %; the last row is $\Delta_{\text{O-C}}$ in percentage points. Guinet: n = 1,000 on the same task domains as VERGE; VERGE: n = 2,854.

	Retriever	Gemma2-9B	Llama3.1-8B	Ministral-8B	Qwen3.5-9B	Average
No verif.	Closed	54.4	54.2	54.2	57.0	54.9
	Oracle	64.7	62.3	61.8	67.8	64.2
	Gap	10.3	8.1	7.6	10.8	**9.2**
Verif. only	Closed	53.8	52.4	50.2	58.0	53.6
	Oracle	70.3	65.0	66.6	71.6	68.4
	Gap	16.6	12.6	16.4	13.6	**14.8**
VERGE	Closed	53.1	52.6	49.9	57.1	53.2
	Oracle	70.3	65.2	66.5	71.1	68.3
	Gap	17.2	12.6	16.6	14.0	**15.1**

Table 16: RAG performance on the ablation study on "no verification", "verification only" and "VERGE" settings across four models (Gemma2-9B, Llama3.1-8B, Ministral-8B, Qwen3.5-9B) and two retrieval configurations (Closed-book, Oracle) with the chunk size of 4000. Values represent accuracy in %. The "Gap" row indicates the difference in accuracy between the Closed-book and Oracle settings for each model and setting, with the average gap calculated across all models.

Metric	Refined	Non-Refined
MH validity	94.67%	95.00%
QA integrity	96.67%	97.00%
Distractor quality	3.51 ($\pm$ 0.60)	3.22 ($\pm$ 0.46)

Table 17: Evaluation of Refined vs. Non-Refined Questions (n=100 per dataset). Standard deviations are reported for the distractor quality metric.

Condition	MH	QA	Both	Dist.
No verif.	0.873	0.705	0.825	0.710
Verif. only	0.850	0.807	0.835	0.783
VERGE	0.860	0.813	0.864	0.788

Table 18: Inter-annotator agreement for the verification ablation: Fleiss' κ for MH, QA and Both; Krippendorff's α for distractor quality (Dist.).

overlap analysis. Table 22 presents the stratified chunk-removal ablation results. Table 23 presents the results of cross-model stress test on GPT-4o and GPT-4o-mini.

K Open-ended Question Answer Generation

K.1 Open-ended Question Answer Generation Prompt

Multi-hop Open-ended Question Generation

Objective:
Generate challenging open-ended questions requiring complex multi-hop reasoning across multiple documentary sources for comprehensive knowledge assessment.

Fundamental Requirements:

1. Questions must necessitate synthesis across **all** provided documentary chunks
2. Answer derivation must be impossible using only subset of chunks
3. Reference answers must demonstrate explicit multi-hop reasoning chains
4. Questions should elicit analytical, comparative, or causal reasoning

Targeted Question Categories:

- **Analytical**: "Analyze the relationship between X and Y based on provided information"
- **Comparative**: "Compare and contrast approaches described in documentation"
- **Causal**: "Explain how X leads to Y, considering all mentioned factors"
- **Synthesis**: "Construct an argument for/against based on all provided information"
- **Evaluative**: "Evaluate effectiveness of X considering mentioned constraints"

Multi-hop Reasoning Patterns:

1. **Chain Reasoning**: Information(Chunk A) $\rightarrow$ Understanding(B) $\rightarrow$ Explanation(C)
2. **Convergent Reasoning**: Multiple chunks provide complementary aspects converging to unified conclusion

	Gemma2-9B	Llama3.1-8B	Ministral-8B	Qwen3.5-9B
One Chunk removed	66.3 (-4.7%)	62.7 (-3.2%)	61.8 (-6.4%)	64.5 (-7.6%)
Single Best Chunk	61.6 (-11.6%)	59.0 (-9.0%)	56.2 (-14.8%)	61.1 (-12.4%)

Table 19: Ablation study results on multi-hop questions. The table shows the accuracy (%) of RAG systems when one chunk is removed from the retrieved set and when only one chunk is provided, compared to the full multi-hop setting. The percentage decrease in accuracy from the full multi-hop setting is indicated in parentheses.

Metric	Value
Mean max F1 (best chunk)	0.200
Median max F1	0.174
Single-best-chunk accuracy	52.1%
Lexical-Oracle gap	47.9%

Table 20: Lexical Overlap Analysis

Quartile	F1 Range	Accuracy
Q1 (lowest overlap)	[0.00, 0.13]	38.5%
Q2	[0.13, 0.17]	41.1%
Q3	[0.17, 0.23]	53.4%
Q4 (highest overlap)	[0.23, 0.96]	75.7%

Table 21: Accuracy by Lexical Overlap Quartile

3. **Comparative Integration**: Contrasting information across chunks to derive insights

4. **Conditional Dependencies**: If-then relationships spanning multiple documentary sources

Input Parameters:
Domain: {task_domain}
Documentation (all chunks mandatory for utilization):
{documentation}

Output Format Specification:
`Question`: [Complex open-ended question requiring multi-hop reasoning]
`Reference Answer`: [Comprehensive answer demonstrating information synthesis from all chunks, minimum 3-5 sentences]
`Reasoning Steps`: [Key reasoning process steps]

Quality Criteria:

- Requires deep understanding and synthesis of all chunks
- Contains clear, comprehensive reference answer
- Prevents shortcut solutions or partial information usage
- Encourages sophisticated reasoning processes

Generate only specified output format

K.2 Open-ended Question Answer Verification Prompt

Multi-hop Reasoning Verification Framework

Objective:
Perform comprehensive verification of multi-hop reasoning requirements for open-ended questions through systematic analysis.

Verification Dimensions:

1. Multi-hop Reasoning Verification

- Determine necessity of cross-chunk information synthesis
- Identify all chunks required for complete answer formulation
- Detect potential reasoning shortcuts bypassing multi-hop requirements

2. Answer Completeness Assessment

- Verify comprehensive synthesis of required information
- Validate explicit multi-hop reasoning demonstration
- Ensure answer non-derivability from chunk subsets

3. Chunk Utilization Analysis

- Confirm essentiality of all provided chunks
- Identify insufficiently integrated chunks
- Assess cross-chunk integration depth

Condition	Low similarity				High similarity			
	Gemma	Llama	Ministral	Qwen	Gemma	Llama	Ministral	Qwen
Full (baseline)	62.8	59.1	57.8	65.1	63.1	60.3	62.2	66.2
One Chunk removed	61.7 (-1.7%)	56.3 (-4.7%)	55.2 (-4.5%)	61.8 (-5.1%)	62.0 (-1.7%)	57.8 (-4.1%)	51.9 (-16.6%)	65.6 (-1.0%)
Single Best Chunk	60.7 (-3.3%)	56.8 (-4.0%)	53.1 (-8.2%)	58.6 (-10.1%)	61.3 (-2.9%)	58.9 (-2.3%)	53.0 (-14.9%)	63.2 (-4.6%)

Table 22: Stratified chunk-removal ablation. Accuracy (%) split by inter-chunk similarity stratum. Percentage change from the full-chunk baseline within each stratum is shown in parentheses.

	GPT-4o	**GPT-4o-mini**
Closed	54.7%	51.0%
Single Best Chunk	60.3%	53.0%
One Chunk Removed	63.3%	60.0%
Oracle	76.0%	69.3%

Table 23: Cross-Model Stress Test: Accuracy under Partial Context Conditions for GPT-4o and GPT-4o-mini.

4. Improvement Recommendations

- Multi-hop failure: Suggest comprehensive chunk incorporation strategies
- Answer incompleteness: Recommend completeness-enhancing details
- Chunk underutilization: Propose integration enhancement methods

Input Variables:
Question to Verify: {question}
Reference Answer: {reference_answer}
Source Documents: {chunk_text}

Output JSON Schema:

```
{
  "required_chunks": [list of essential chunk indices],
  "reasoning": {
    "multi_hop_required": boolean,
    "answer_completeness": boolean,
    "shortcut_possible": boolean,
    "unused_chunks": [list of underutilized chunk indices],
    "improvement_suggestions": string (mandatory)
  },
  "confidence": integer (1-5 scale)
}
```

Critical Constraints:
Ensure question non-answerability through:

- Partial chunk utilization
- General knowledge without documentary evidence
- Surface-level reading without synthesis

Output JSON response exclusively

K.3 Open-ended Question Answer Regeneration Prompt

Multi-hop Question Regeneration Protocol

Objective:
Regenerate open-ended questions to satisfy comprehensive multi-hop reasoning requirements based on verification feedback.

Regeneration Context:
Original Question: {original_question}
Original Answer: {original_answer}

Verification Feedback:

- **Unused Chunks Requiring Incorporation**: {unused_chunks}
- **Specific Improvement Requirements**: {suggestions}

Regeneration Requirements:

1. Mandate information utilization from **all** provided chunks
2. Prohibit answerability using chunk subsets
3. Ensure clear, comprehensive reference answer formulation
4. Demonstrate genuine multi-hop reasoning pathways

5. Eliminate shortcuts and surface-level reasoning possibilities

Input Parameters:
Domain: {task_domain}
Documentation:
{chunk_text}

Output Format Specification:
`Question`: [Improved question requiring all chunks]
`Reference Answer`: [Comprehensive answer synthesizing all chunks, 3-5 sentences]
`Improvements Made`: [Brief description of feedback incorporation]

Quality Assurance Criteria:

- Complete feedback incorporation
- All chunk utilization verification
- Multi-hop reasoning pathway clarity
- Answer comprehensiveness validation

Generate improved question addressing all feedback

K.4 Open-ended Question Answer Example

K.4.1 Example GovReport Question

Question:
Analyze the legal challenges faced by students accused of sexual misconduct in public universities, considering the due process concerns outlined in the provided documentation. How do these challenges relate to the university's responsibility to ensure fairness and transparency in disciplinary proceedings?

Reference Answer: Students accused of sexual misconduct in public universities face significant legal challenges related to due process violations. The documentation highlights several key issues, including inadequate notice of charges, limited opportunity to prepare a defense, and unfair review processes for appeals. Courts have recognized that universities must provide students with timely and comprehensive notice of allegations, a meaningful opportunity to respond, and access to evidence used against them. Furthermore, universities must ensure that appeal processes are fair and transparent, providing students with adequate notice of new evidence, the opportunity to respond, and a chance to present their case orally. Failure to uphold these principles can result in a deprivation of a student's liberty interest without due process of law.

K.4.2 Example HotpotQA Question

Question:
Analyze the evolution of loan modification programs in the context of the housing crisis, comparing and contrasting the approaches taken by the state court RMFM program, the Middle District Bankruptcy Court's LMM program, and the Fannie Mae/Freddie Mac SMP. What factors contributed to the success or failure of each program?

Reference Answer: The housing crisis spurred various loan modification programs. The state court RMFM program, while aiming to provide alternatives to foreclosure, ultimately failed due to a lack of enforcement and cooperation from lenders. In contrast, the Middle District Bankruptcy Court's LMM program, with its stricter requirements and focus on mediation, achieved higher success rates. The Fannie Mae/Freddie Mac SMP, a fast-track program, aimed to reduce monthly mortgage payments to 38% of gross income, offering a more standardized approach to affordability. The success of the LMM program highlights the importance of strong enforcement mechanisms and lender accountability, while the SMP's standardized approach demonstrates a different strategy for addressing affordability challenges.

K.4.3 Example MultiFieldQA Question

Question:
Analyze the authors' approach to dialect identification in the context of the challenges and advancements discussed in the provided documentation. How does their two-stage pipeline address the complexities of distinguishing between similar languages and dialects, and what evidence do they provide to support its effectiveness?

Reference Answer: The authors propose a two-stage pipeline for dialect identification, consisting

of a language identification (LID) model followed by language-specific dialect classification models. This approach addresses the challenge of distinguishing between similar languages and dialects by first accurately identifying the language of the input text, then leveraging language-specific models for more precise dialect classification. They demonstrate the effectiveness of their pipeline by achieving top performance in both tracks of the VarDial 2023 DSL-TL shared task, outperforming other participants and previous works. Their results, including a 58.54% accuracy in Track-1 and 85.61% in Track-2, provide strong evidence for the superiority of their two-stage approach.

K.4.4 Example SecFilings Question

Question:
How does PetroCorp's approach to asset maintenance and drilling strategies, as described in their forward-looking statements, aim to mitigate the inherent volatility of oil and gas markets, particularly the fluctuations in prices, as outlined in their risk factors?

Reference Answer: PetroCorp acknowledges the volatile nature of the oil and gas market, recognizing that prices are susceptible to fluctuations due to supply and demand, market uncertainty, and external factors beyond their control. To mitigate these risks, the company emphasizes a two-pronged approach. Firstly, PetroCorp prioritizes maintaining its property, plant, and equipment in good operating condition, as stated in their forward-looking statements. This proactive asset management strategy aims to minimize operational disruptions and ensure efficient production, thereby reducing the impact of price volatility on their operations. Secondly, PetroCorp emphasizes careful planning and execution of drilling activities, aiming to optimize production while acknowledging the uncertainties inherent in estimating reserves and production rates. This cautious approach to exploration and production helps to manage the risks associated with fluctuating oil and gas prices. By combining these strategies, PetroCorp seeks to navigate the volatile oil and gas market while maximizing its operational efficiency and profitability.

K.4.5 Example 2WikiMultiHopQA Question

Question:
Analyze the evolution of Wheatstone's contributions to communication technology, considering both his successes and the controversies surrounding his work. How did his inventions impact the development of telegraphy and public perception of technological advancements?

Reference Answer: Wheatstone's contributions to communication technology were multifaceted, ranging from the chronoscope for precise time measurement to the groundbreaking invention of the electric telegraph. While he faced accusations of plagiarism from Alexander Bain, his work on the telegraph, particularly the single-needle instrument, revolutionized communication speed and efficiency. The public's embrace of the telegraph, exemplified by its use in capturing John Tawell, solidified its importance. However, the controversy surrounding his collaboration with Cooke highlighted the complexities of invention and recognition in a rapidly evolving technological landscape.