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

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

Abstract

Despite the rapid growth of retrieval-augmented generation (RAG) systems in industry, existing evaluation datasets inadequately assess multi-hop reasoning capabilities when deployed on custom enterprise knowledge bases, creating a critical evaluation gap between public benchmarks and real-world performance. We propose VERGE (VERification-enhanced GEneration), a two-fold RAG evaluation generation pipeline that (1) employs a Large Language Model (LLM) based verifier to enforce the logical multi-hop reasoning and question-answer (QA) integrity criteria during the question generation process and (2) iteratively refines any questions that failed those criteria. Across 11,052 candidate questions spanning five domains, VERGE filters or refines low-quality items, yielding 2,854 verified multi-hop questions. Human evaluation confirms high verifier reliability (Cohen's $\kappa = 0.903$) and rates the VERGE-generated dataset significantly higher on QA integrity and distractor quality than the existing method ($p < 0.001$). We further propose a hierarchical taxonomy of RAG failure modes, dividing them into Information Processing and Knowledge Boundary errors. Our analysis reveals that the latter, particularly context utilisation failures dominate across all LLMs.

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.

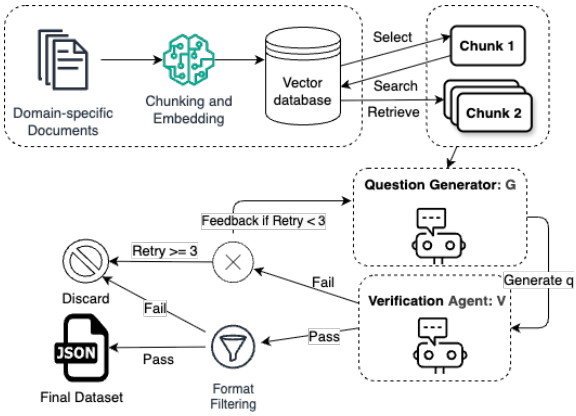

Figure 1: VERGE dataset generation process.

We introduce VERGE (VERification-enhanced GEneration), a novel methodology for generating task-specific multi-hop datasets that evaluates RAG systems' complex reasoning capabilities. Our primary technical contribution is a verification agent that performs dual functions: (1) a LLM-based verifier to enforce the genuine multi-hop reasoning and question-answer (QA) integrity criteria during the question generation process and (2) iteratively refines any questions that failed those criteria. Unlike existing approaches that rely on post-hoc statistical optimisation (Guinet et al., 2024), VERGE incorporates verification directly into the generation process, enabling robust quality control at scale without human intervention.

Building on this methodological foundation, we propose a hierarchical error taxonomy derived from observed RAG failure patterns in multi-hop reasoning scenarios. This taxonomy categorises errors into Information Processing Errors and Knowledge Boundary Errors. Through experiments, we demonstrate that knowledge boundary errors, particularly context utilisation failures, dominate across all evaluated models.

In summary, this paper makes the following contributions:

- We introduce VERGE dataset generation method for evaluating task-specific RAG systems, validated through comprehensive human assessment and experiments across five diverse task domains.
- We propose a hierarchical error taxonomy specifically designed for multi-hop RAG system analysis.

2 Related Literature

A range of benchmark datasets have been proposed to evaluate multi-hop reasoning in question answering. Early datasets such as HotpotQA (Yang et al., 2018), MultiRC (Khashabi et al., 2018), and 2WikiMultihopQA (Ho et al., 2020) established foundational frameworks for evaluating multi-hop reasoning over multiple documents. 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 constructed from multiple news articles specifically evaluating multi-hop queries in RAG contexts. While MultiHop-RAG addresses the need for multi-hop evaluation in RAG systems, it functions as a benchmark dataset rather than a methodological framework for generating task-specific evaluation instruments. Furthermore, these benchmarks typically rely on predetermined document collections, which limits their applicability to domain-specific RAG evaluation scenarios where practitioners need to assess RAG performance on specialised knowledge bases.

In terms of RAG evaluation dataset generation methods, Es et al. (2024) introduced RAGAS for LLM-based open-ended QA generation. However, it does not incorporate any verification mechanism during the generation process, potentially compromising the quality and validity of the resulting assessments. Subsequent studies explored more structured LLM-based generation to better control reasoning properties (Zhu et al., 2025). Several works impose compositional or multi-step structures during generation (Shen et al., 2025; Lee et al., 2025b), while others introduce explicit difficulty or robustness dimensions (Gong et al., 2025; Lee et al., 2025a; Zeng et al., 2025). Although these frameworks provide valuable diagnostic signals, they typically rely on post-hoc filtering or aggregate difficulty statistics, and do not explicitly enforce - at the level of individual instances - that all retrieved evidence is necessary to answer the question. Closest to our work, Guinet et al. (2024) proposed an approach through an LLM-based multiple-choice exam generation framework for task-specific RAG assessment. Their methodology employs post-hoc statistical optimisation, where RAG systems first solve generated questions, then Item Response Theory (IRT) is applied to evaluate exam quality and iteratively refine the question set. While this represents methodological progress, it postpones verification until after the generation process, potentially introducing computational inefficiency and limiting the dataset quality of the initial question set. In contrast, VERGE targets distractor-rich multiple-choice question generation for multi-hop RAG evaluation, integrating verification into the generation process. Although prior research has explored refinement mechanisms of RAG components during the response generation (Shao et al., 2023; Asai et al., 2024; Zhou et al., 2024; Shi et al., 2024), our method applies self-reflection to RAG evaluation dataset generation. Consequently, we primarily compare VERGE against Guinet et al. (2024), which represents the closest prior work in multiple-choice RAG assessment.

The RAG failure pattern categorisation remains an emerging area of research with limited number of literature. Simon et al. (2024) manually reviewed the RAG failure modes and categorised them qualitatively. Li et al. (2024) and Xu et al. (2024) categorised the RAG error patterns by leveraging LLMs. These preliminary taxonomies provide valuable insights. Nevertheless, these taxonomies predominantly focus on single-hop reasoning contexts and do not capture the unique failure patterns of multi-hop reasoning. Following those existing works, we propose an error taxonomy for multi-hop RAG. Through manual analysis of multi-hop RAG failure patterns, we identify distinct error categories related to information integration and knowledge boundary errors as a primary failure pattern.

3 VERGE framework

3.1 Theoretical Framework

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 with a maximum size of 4,000 characters. A multi-hop question q requires information synthesis across a subset of chunks $C_q \subset C$, where $2 \leq |C_q| \leq 5$. C_q is constructed by first randomly selecting a chunk c_1 from C, then retrieving the top-k relevant chunks based on semantic similarity. Formally, we define a multi-hop question as:

$$q = \langle question, \{a_1, a_2, a_3, a_4\}, a^*, C_q \rangle, \quad (1)$$

where question represents the question text, $\{a_1, a_2, a_3, a_4\}$ are candidate answers, a^* is the correct answer, and C_q is the set of chunks required to determine a^*.

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^*. The existing method (Guinet et al., 2024) struggles to generate effective distractors and implements a separate independent process to generate high-quality distractors. However, we approach this challenge with extensive prompt engineering and integrate distractor engineering directly within the generation workflow. 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). Our approach demonstrates that the well-constructed prompt engineering can outperform the multi-iteration distractor generation.

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 ensures generated questions meet multi-hop requirements. During the question generation process, the verification agent verifies the question following two criteria: (1) **Complete Multi-Hop**: $V_c(q, C_q) \rightarrow \{0, 1\}$ ensures that 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 enforces an operational criterion reflected by performance degradation under partial context. (2) **QA Integrity**: $V_v(q, C_q) \rightarrow \{0, 1\}$ confirms 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 format of equation 1, q will be discarded.

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

Methodology	Timing	Feedback	Criteria
Es et al. (2024)	None	None	None
Guinet et al. (2024)	Post	None	IRT model
VERGE (Ours)	During	$\leq$ 2 iter	Complete MH and QA integrity

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); *Feedback*: number of iterative refinement cycles (iter = iterations); *Criteria*: evaluation metrics used.

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**	**51,943**	**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.

3.4 VERGE Dataset

For the verification agent, we use Gemma2-9B (Team et al., 2024), tests with Llama 3.2-3B (Meta, 2024) and Ministral-8B (Mistral AI, 2024) showed $< 2\%$ variance in pass rates. Relying on one LLM may cause an unwanted bias such as self-favouring bias (Zheng et al., 2023; Panickssery et al., 2024; Wang et al., 2024). To minimise the effect of such

bias, we employ three diverse question generator models: Llama3.2-3B, Gemma2-9B, and Ministral-8B of varying sizes and architectures.

Our dataset consists of 2,854 multi-hop questions across six different task domains: $T = \{t_{\text{gov}}, t_{\text{hot}}, t_{\text{multi}}, t_{\text{wiki}}, t_{\text{covid}}, t_{\text{sec}}\}$. We employ four datasets, GovReport (Huang et al., 2021), HotpotQA (Yang et al., 2018), MultiFieldQA (Bai et al., 2023), and 2WikiMultihopQA (Ho et al., 2020) from LongBench datasets (Bai et al., 2023) 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. The number of chunks in a vector database varies across domains, from 270 chunks in t_{covid} to 36,834 in t_{sec}, with an average chunk size of 338 words. 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 verification agent integrated into the question generation process enables guided generative refinement and distinguishes our approach from previous work that relies primarily on statistical filtering or post-hoc optimisation.

Our dataset exhibits diverse reasoning requirements. Building on the question type taxonomy proposed by Tang and Yang (2024), we classify questions using the Gemma2-9B model into three non-mutually exclusive categories: inference, comparison, and temporal. 54.8% of questions in our dataset are categorised as inference reasoning, 25.2% combining inference and comparison, and 9.8% integrating inference, comparison and temporal reasoning types. In terms of reasoning complexity, 52.4% of questions require synthesising information across three or more chunks to arrive at the correct answer.

3.5 Dataset Evaluation Method

To empirically assess the quality of the VERGE-generated dataset, we conduct a human annotation study. This study aims: (1) to manually validate the correctness of the verification agent's assessment and (2) to evaluate the overall quality, specifically QA integrity and distractor quality of VERGE dataset against the Guinet et al. (2024) dataset. Guinet dataset was generated employing the claude-3-5-sonnet@20240620 model (Anthropic, 2024) by using their publicly released codebase on the same task domains and chunks as the VERGE dataset.

We engage three independent machine learning practitioner annotators to review three datasets: 100 samples from VERGE-generated dataset, 100 samples that failed either or both VERGE verification criteria (Complete Multi-Hop and QA Integrity), and 100 samples from the Guinet dataset.

Each annotator evaluates 300 QA pairs using the following rubric:

- Verification Criteria Adherence: For each QA pair, assess whether the question satisfies (a) Complete Multi-Hop and (b) QA Integrity.
- Distractor Quality: Rate the quality of the three distractors on a 1 (very poor) to 5 (excellent) scale based on relevance, plausibility, and difficulty.

We employ multiple evaluation metrics to analyse the results. Cohen's Kappa (Cohen, 1960) is used to assess the agreement between the verification agent and annotators based on binary pass/fail judgments. For inter-annotator agreement, we calculate Fleiss' Kappa (Fleiss, 1971) to evaluate the consistency across all three annotators. We apply the Chi-squared test to compare the proportion of QA pairs satisfying the verification criteria across the VERGE dataset and the Guinet dataset. We also use the Mann-Whitney U test (Mann and Whitney, 1947) to assess differences in distractor quality scores between the VERGE dataset and the Guinet dataset. All annotators evaluate an identical set of QA pairs presented in randomised order to reduce potential order effects.

3.6 RAG Experiment Configuration

We employ four open language models; Llama3.1-8B (Dubey et al., 2024), Ministral-8B, Gemma2-9B and Qwen3.5-9B (Yang et al., 2025) 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:** Question and answer candidates are given to the LLM without any retrieved chunks - serves as a lower bound of the experiment, **Oracle:** Question, answer candidates and the ground truth chunks are given to the LLM - serves as an upper bound of the experiment, **Dense:** Utilising bge-large-en-v1.5 (Xiao et al., 2023) 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) (Wei et al., 2022) We adapt prompts for the method from Zhuang et al. (2024). As our dataset is a multiple-choice format, we employ point-wise accuracy metric to evaluate the performance of the RAG pipelines.

To assess whether in-process verification translates into amore discriminative evaluation instrument, we additionally conduct a comparative analysis of RAG performance on the VERGE dataset against the Guinet dataset (Guinet et al., 2024). This controlled setup isolates the effect of dataset construction methodology on benchmark discriminativeness, holding all other experimental variables constant.

3.7 Error Category Analysis

Through manual review of failed RAG cases, we implement a hierarchical error taxonomy framework to analyse RAG systems' failure patterns on multi-hop reasoning tasks. The taxonomy comprises two primary categories: **Information Processing Errors**, emcompassing **integration failures** and **reasoning failures**, and **Knowledge Boundary Errors**, encompassing **context utilisation** errors and **knowledge scope violations**. We employ Gemma2-9B as our error classifier in this experiment. The classifier receives a QA pair and the model's reasoning chain, and the source documents, and assigns a primary category, a subcategory, and a natural-language justification. The full classification prompt is provided in Appendix D.3.

To validate the reliability of the automated classifier, we conduct a human annotation study. Three independent annotators label 100 stratified samples across the two primary error categories, sampled to reflect the empirical distribution in the data (80% Knowledge Boundary, 20% Information Processing). Annotators assign a primary category and subcategory to each instance based on the same inputs provided to the classifier (QA pair, model reasoning, and source documents). Inter-annotator agreement is assessed by Fleiss' Kappa and per-category precision and recall. We additionally evaluate Claude Sonnet 4.5 as an independent reference classifier under identical conditions to assess whether the validation generalises beyond the Gemma2-9B model family.

	Complete MH	QA Integrity	Both
VERGE-Passed	95.0%	97.0%	93.3%
VERGE-Failed	4.0%	65.7%	3.0%
Cohen's Kappa	0.903	0.865	0.903
Fleiss' Kappa	0.860	0.813	0.864

Table 3: Human validation of VERGE verification criteria. Values represent the percentage of questions rated as passing by human annotators across 300 annotations per dataset. Cohen's Kappa indicates agreement between VERGE agent and human assessments. Fleiss' Kappa indicates inter-annotator agreement.

4 Experiment Results

4.1 Dataset Evaluation

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 significant disparity confirms the discriminative strength of the VERGE verification agent in filtering 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, which indicates near-perfect agreement. Fleiss' kappa scores, which reflect inter-annotator consistency across all three raters, were also high: 0.860 for Complete Multi-Hop, 0.813 for QA Integrity, and 0.864 for overall pass/fail decisions. These results provide strong evidence that the VERGE verification mechanism reliably aligns with human judgment.

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 (Pearson, 1900) (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, $p < 0.001$).

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 (higher is better) and reported with standard deviation. Statistical significance is determined via a Chi-squared test (QA Integrity) and Mann-Whitney U test (Distractor Quality).

This performance is particularly notable given that the Guinet dataset was generated using a more capable model than those employed in VERGE. These results underscore the effectiveness of the VERGE framework compared to the existing evaluation dataset generation method: by integrating an LLM-based verification agent during generation and applying iterative refinement, it substantially enhances both QA integrity and distractor quality.

4.2 RAG Performance

The experiment results in Figure 2 reveal several findings about RAG system optimisation patterns. First, almost all RAG configurations outperform the closed-book baseline and underperform the Oracle setting, which confirms the utility of VERGE dataset as a robust benchmark for multi-hop reasoning tasks. Second, CoT prompting enhances performance compared to standard prompting only when the chunk size is sufficiently large. Under the sparse retriever, Gemma2-9B achieves higher accuracies (up to 65.3%) with the standard prompting compared to CoT prompting (up to 63.3%) for chunk size of up to 2,048, yet CoT prompting outperforms standard prompting (65.6% vs 63.7%) at chunk size of 4,000. We conjecture that more capable models are better able to process longer, semantically nuanced chunks and synthesise information from them.

The results of RAG performance comparison on the VERGE and Guinet dataset reveals that the discriminative gap (Oracle-Closed) is consistently larger on VERGE by 5.2-7.3 percentage points across all models, which demonstrates the VERGE method yields a measureably more discriminative evaluation instrument. Moreover, closed-book accuracy on the Guinet dataset substantially exceeds that on VERGE, indicating that a large proportion of Guinet questions are answerable from parametric knowledge alone or distractors are obvious. The full results of the both experiments are reported in Appendix I.

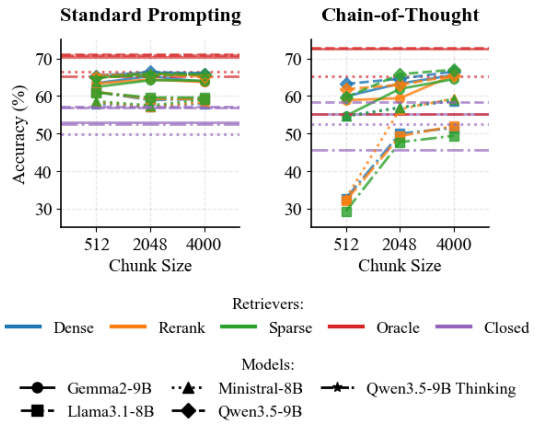

Figure 2: Accuracy in %. Results across four LLMs, three chunk sizes (512, 2,048, 4,000 characters), three retrievers, and two prompting strategies (Standard and Chain-of-Thought), plus closed-book and Oracle settings. For Qwen3.5-9B, Thinking denotes the thinking-enabled variant with standard prompting.

4.3 Error Category Analysis

Prior to reporting the error distribution, we establish the reliability of the Gemma2-9B classifier. Three human annotators show strong agreement among themselves (Fleiss' Kappa = 0.77) and the majority vote among annotators achieves Cohen's Kappa = 0.97 against Gemma2-9B classifier. Per-category analysis reveals no structural bias (Information Processing: Precision = 0.95, Recall = 1.0; Knowledge Boundary: Precision = 1.0, Recall = 0.99). Claude Sonnet 4.5, evaluated as an independent reference, shows strong alignment with Gemma2-9B (Cohen's Kappa = 0.80). These results establish that Gemma2-9B constitutes a reliable proxy for human judgment. The full results of the error category evaluation are in Appendix L.3.

The error category distribution reveals that Knowledge boundary errors significantly outweigh information processing errors across all models, accounting for approximately 80% of total errors. Among these, context utilisation failures are the dominant error type, where models fail to properly leverage information across multiple documents. Representative examples of each error category and subcategory, including classification justifications are provided in the supplementary material. The proportional distribution of errors remains consis-

tent across different model architectures for standard prompt, suggesting these failure patterns represent intrinsic challenges in RAG architectures rather than model-specific limitations. This finding has significant implications for RAG system design, indicating that improvements targeting context utilisation may achieve substantial performance gains across model families.

Regarding the prompting strategies, CoT prompting reduces information processing errors, particularly integration failures, while the proportion of reasoning errors remains relatively stable for Llama3.1-8B and Ministral-8B models. Conversely, context utilisation errors increase with CoT prompting, especially for Gemma2-9B, rising from 53.8% to 76.2%. This suggests that while CoT enhances reasoning structure, it may lead models to become more selective about incorporated information, potentially overlooking crucial evidence. These patterns explain why even in Oracle settings, models struggle to effectively synthesise information across multiple sources, highlighting context utilisation as the primary challenge in multi-hop RAG systems.

5 Analysis and Discussion

5.1 Ablation of Verifier

The verification agent relies on two criteria to filter and refine the dataset: complete multi-hop and QA integrity. We observe the number of eliminated questions per verification criteria in Table 5. The complete multi-hop criteria demonstrates significantly higher filtering impact, eliminating 72.1% of candidates from the initial dataset. This reflects the difficulty of generating questions that require information synthesis across multiple chunks, even with iterative refinement opportunities. In contrast, the QA integrity criteria filters a comparatively modest 8.2% of questions, suggesting that generating a question with factual consistency is less challenging to maintain than multi-hop reasoning requirements. Despite their asymmetric filtering effects, both criteria demonstrate complementary value in maintaining dataset quality. Figure 3 in appendix F illustrates the refinement process, showcasing how the feedback mechanism guides the incorporation of critical information from an unused chunk, transforming what would otherwise be rejected candidates into a valid multi-hop question. These analyses exhibit both filter and refine capabilities of the verification agent represent an integral part of VERGE framework and differentiate our method to other existing RAG evaluation dataset generation methods.

Criteria	# eliminated	% eliminated
Complete Multi-Hop	7,967	72.1%
QA Integrity	902	8.2%
Format Filtering	245	2.2%

Table 5: Number of eliminated questions per verification criteria and final formatting filtering. Initially there are 11,052 candidate questions and survived 2,854 questions.

In terms of the effectiveness of the refinement process, our analysis reveals that 326 questions, which is 11.4% of the final dataset were successfully refined through verification feedback. Out of 326 refined questions, 260 questions were refined with one iteration and 66 questions required two iterations. While 11.4% of final questions were refined, these are not borderline cases and refinement targets structurally complex questions. Human annotation study, three annotators for 100 sampled QAs, comparing refined and non-refined items reveals that refined questions exhibit statistically higher hop complexity (Mann-Whitney: $p < 10^{-13}$). MH validity and QA integrity show no significant difference ($p = 0.85$ and $p = 0.81$, respectively), confirming that refined questions meet the same quality threshold as first-pass successes despite being structurally more complex. Distractor quality is significantly higher for refined items (Mann-Whitney: $p < 10^{-6}$). The full results can be found in appendix J.

5.2 Ablation of Complete Multi-Hop

The complete multi-hop criterion is evaluated by ablating chunks from the context used to answer each question in the VERGE dataset. Under the MCQ format, removing the single best chunk yields an 11.8% accuracy drop, and removing one chunk from the full context yields a 4.8% drop on average relative to the Oracle setting. While these figures confirm degradation, the MCQ format partially compresses the observed gap: models benefit from a 25% random baseline and can eliminate implausible distractors using partial information, which makes genuine multi-hop dependence. To obtain a format-unconfounded estimate, we re-evaluate 1,125 QA pairs under an open-ended setting by removing answer options and reporting both token level F1 score and LLM-as-a-judge

accuracy. Under the open-ended format, single-chunk conditions yield average F1 and LLM-judge drops of 36.0% and 15.3% respectively, substantially larger than the MCQ-format figures. This confirms that the multi-hop dependency is genuine but was partially masked by the multiple-choice format in the primary experiments. Convergent evidence is provided by a cross-model stress test using GPT-4o and GPT-4o-mini on 300 sampled QA pairs, evaluating under closed-book, single-best-chunk, one-chunk-removed, and Oracle conditions. McNemar's tests confirm statistically significant degradation under all partial-context conditions for both models (all $p < 10^{-4}$), with 9-16 percentage-point gaps between full-context and single-chunk performance for GPT-4o. The monotonic Closed < Single Best < One Less < Oracle order holds across all models, confirming that the multi-hop criterion reflects genuine distributed reasoning rather than verifier-specific limitations.

To further characterise the multi-hop requirement, we analyse the relationship between lexical overlap and performance. We define the theoretical retrieval ceiling as 100%, reflecting that all questions passed QA Integrity verification and are answerable from the provided chunks by construction. We note this is distinct from the empirical model accuracy under Oracle retrieval reported in Table 11, which additionally reflects model-side reasoning limitations. Single-best-chunk accuracy reaches only 52.1%, marginally above the 25% random baseline for a four-choice MCQ. Even in the highest lexically overlapped quartile (Q4), accuracy reaches only 75.7%, confirming that surface-level keyword matching is insufficient to resolve VERGE questions and that multi-hop information synthesis is genuinely required. These results collectively provide converging evidence that VERGE's multi-hop requirement reflects genuine distributed reasoning rather than lexical shortcuts. The full results are presented in appendix K.

5.3 Cost and Generator Tradeoffs

The average wall time per iteration depth and monetary cost of all the generator models along with Claude 4.5 Sonnet (Anthropic, 2025b) and Haiku (Anthropic, 2025a). Across all models, generation completes within 6 - 45 seconds depending on the model size and number of iterations. API usage is similarly modest, costing under $5 per 100 QA pairs for Sonnet and under $2 for Haiku. To characterise the practical tradeoff between pipeline cost and dataset yield, we generated 200 QA pairs using a Claude-based pipeline (Sonnet 4.5 as generator; Haiku 4.5 as verifier) on the same task domains and chunks as the open-weights pipeline. The Claude pipeline achieves a substantially higher acceptance rate (63%) compared to the open-weights pipeline (22%), with a greater proportion of questions passing without any refinement (33% vs. 18%), reflecting higher initial generation quality. Crucially, both pipelines exhibit a monotonic increase in average hop count with refinement depth, demonstrating that the verification-and-refinement mechanism rescues structurally complex questions regardless of the generator model family and size. The full results including the wall time and acceptance rate and iteration breakdown by pipeline are provided in Appendix H.4.

5.4 Open-Ended Question Generation

While our main evaluation focuses on MCQ format, VERGE framework naturally extends to open-ended questions. The open-ended variant maintains our verification mechanism, ensuring multi-hop reasoning requirements while generating comprehensive reference answers. Out of 600 generated open-ended question candidates, 56.0% of QA pairs passed the verification, which is notably higher than the 23.1% acceptance rate for multiple-choice questions. This difference reflects the inherent complexity of distractor generation rather than multi-hop reasoning requirements, highlighting that distractor creation remains a fundamental challenge. The prompt and examples of open-ended questions generated by VERGE are found in the appendix M.

6 Conclusion

We present VERGE that advances the methodological rigour of multi-hop RAG evaluation dataset construction. By integrating an LLM-based verifier during the generation process, VERGE ensures both multi-hop reasoning completeness and QA integrity, enabling robust and scalable dataset curation with minimal human intervention. Our empirical evaluation demonstrates that this in-process verification significantly improves both question and distractor quality compared to the existing method.

VERGE's dual capabilities of discriminative filtering and constructive refinement enable the generation of high-quality multi-hop questions that better reflect real-world task-specific demands.

References

Anthropic. 2024. Introducing the next generation of claude. Accessed: 2026-01-02.

Anthropic. 2025a. Introducing claude haiku 4.5. Accessed: 2026-01-02.

Anthropic. 2025b. Introducing claude sonnet 4.5. Accessed: 2026-01-02.

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, Hong Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *ArXiv*, abs/2308.14508.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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, Kamilė Stankevičiūtė, 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. *arXiv preprint arXiv:2502.20377*.

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.

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.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

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 stochasti-

cally larger than the other. *The annals of mathematical statistics*, pages 50–60.

Meta. 2024. Model cards & prompt formats llama 3.2. Accessed: 2026-01-02.

Mistral AI. 2024. Introducing Ministraux: The world's best edge models. Accessed: 2026-01-02.

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.

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.

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.

Sebastian Simon, Alina Mailach, Johannes Dorn, and Norbert Siegmund. 2024. A methodology for evaluating rag systems: A case study on configuration dependency validation. *arXiv preprint arXiv:2410.08801*.

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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: 2026-05-09.

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. *ArXiv*, abs/2002.10957.

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*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yunqi Xu, Tianchi Cai, Jiyan Jiang, and Xierui Song. 2024. Face4rag: Factual consistency evaluation for retrieval augmented generation in chinese. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 6083–6094, New York, NY, USA. Association for Computing Machinery.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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*.

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-favoring 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 generation prompts and criteria to support transparency and auditing. We view VERGE as complementary to, rather than a replacement for, diverse evaluation methodologies.

A.1 Licensing

The source corpora are derived from LongBench (MIT License; Bai et al. (2023)), 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) and Claude models (Consumer 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.

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 Error Categorisation Prompt

Multi-Hop Question Answering Error Analysis Methodology Objective:
You are an expert evaluator analyzing errors in multi-hop question answering systems. Your task is to perform a detailed error analysis using a hierarchical classification system.

Primary Error Categories:

1. **Information Processing Errors**
   A. **Integration Failures**
      - Incorrect combination of retrieved facts
      - Failure to establish relationships between facts
      - Missing key connections between pieces of information
   B. **Reasoning Failures**
      - Invalid logical inference steps
      - Incorrect causality assumptions
      - Faulty deductive or inductive reasoning
2. **Knowledge Boundary Errors**
   A. **Context Utilization**
      - Missing or overlooking relevant information
      - Using irrelevant or misleading information
      - Failure to identify key supporting evidence
   B. **Knowledge Scope**
      - Making assumptions beyond given information
      - Introducing external knowledge inappropriately
      - Misinterpreting information boundaries

Analysis Instructions:

1. Identify error chains by tracing the progression of mistakes
2. If none of the categories are the right category, return "other"
3. For each error identified, provide:
   - Evidence from the exam taker reasoning
   - Specific text from source documents
   - Explanation of how the error led to the incorrect answer

Output Format:
Return strictly only JSON format as specified below and do not return other explanations.

JSON Schema:

- `error_chain`: Dictionary containing:
  - `primary_error`: Dictionary with:
    - `category`: String (Main category from hierarchy)

        - subcategory: String (Specific error type either A or B)
        - confidence: Integer (1-5 scale)
        - evidence: Dictionary containing:
            - exam_taker_reasoning: String (Relevant quote)
            - source_document: String (Relevant quote)
            - explanation: String
    - contributing_errors: Array of dictionaries (Same structure as primary_error)

- quantitative_indicators: Dictionary containing:
    - information_overlap: Float (Semantic similarity with source)
    - reasoning_steps_identified: Integer
    - external_knowledge_ratio: Float (Proportion of unsupported claims)

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

Fundamental Constraints:

- Generate only JSON response following the specified schema
- Maintain rigorous hierarchical error classification
- Provide comprehensive evidence-based analysis
- Ensure traceability of error progression chains

Perform comprehensive multi-hop error analysis

D.4 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.5 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 are arbitrary selected generated question per task domain in the dataset.

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.2.1 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.2.2 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.2.3 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	Llama3-3B	Ministral-8B
Generated	3,670	3,714	3,670
QAs	1,199	822	1,007
Success	32.7	22.1	27.4

Table 6: 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.

F Verification Process Survival Rate

Table 6 presents the survival rates of generated question-answer pairs across the three LLMs used for generation.

However, the number of surviving questions varies across task domains, with the lowest survival rate observed in the HotpotQA dataset (16.7%). This discrepancy is likely due to the inherent complexity of the HotpotQA dataset, which requires synthesising information across multiple documents to answer questions correctly. The GovReport dataset, on the other hand, has the highest survival rate (32.9%), indicating that the questions generated in this domain are more straightforward and less reliant on multi-hop reasoning.

G Detailed Dataset Characteristics

Table 7 presents the distribution of reasoning types in our dataset. Following Tang and Yang (2024)'s taxonomy, we categorise questions based on their reasoning requirements into three fundamental types, yet given the complexity of multi-hop questions, we consider their combinations as well. The majority of questions focus on inference reasoning in our dataset, requiring the derivation of logical conclusions by synthesising information across retrieved chunks, C_q. The most common complex pattern combines inference and comparison reasoning, while 9.8% of questions integrate all three reasoning types.

Beyond reasoning types, our benchmark also varies in information synthesis requirements. While 47.6% of questions require synthesising information from two chunks, a substantial portion demands more complex integration: 19.4% require three chunks, 15.8% require four chunks, and 17.1% require five chunks. This distribution ensures our benchmark evaluates different complex information synthesis capabilities.

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.

Detailed Reasoning Type	# Questions	Percentage
Single Type		
Inference	1,255	55.9%
Comparison	56	2.5%
Temporal	3	0.1%
Two Types		
Inference + Comparison	569	25.2%
Inference + Temporal	61	2.7%
Comparison + Temporal	4	0.2%
Three Types		
Inference + Comparison + Temporal	211	9.3%
Other	99	4.4%
Total	2,258	100.0%

Table 7: Detailed distribution of reasoning types in our benchmark dataset, including all combination patterns.

H RAG Implementation Details for Reproducibility

H.1 Model Sources

We list all the model versions we used in our experiments below in the table 8.

H.2 Model Parameters

For reproducibility purpose, we list language model parameters in the table 9.

H.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.

H.4 Cost and Generator Tradeoffs

Table 10 presents the average wall time per question-answer pair generation across different verifier depths and models.

I RAG results details

Table 11 demonstrates the full results of the RAG experiment with its number of questions in the dataset. Due to the parsing errors, each reported result has different total number of questions.

Table 12 demonstrates the RAG results only with CoT prompting strategy with its number of questions in the dataset.

Table 13 presents the RAG performance comparison between VERGE and Guinet dataset across three models (Llama3.1-8B, Ministral-8B, Gemma2-9B) and four retrieval configurations (Closed-book, Sparse, Dense, Oracle). Values represent accuracy in %.

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
Claude 3.5 Sonnet	2024-06	claude-3-5-sonnet@20240620
Claude 4.5 Sonnet	2025-09	claude-sonnet-4-5
Claude 4.5 Haiku	2025-09	claude-haiku-4-5
Qwen3.5-9B	2026-03	unsloth/Qwen3.5-9B-GGUF (Qwen3.5-9B-Q4_K_M.gguf)

Table 8: List of LLMs used in our experiments

Model	Temperature	top-p	Context window
Gemma2-9B	0.0	0.9	8,192
Llama3.1-8B	0.0	0.9	8,192
Llama3.2-3B	0.0	0.9	8,192
Ministral-8B	0.0	0.9	8,192
GPT-4o	0.0	0.9	128,000
GPT-4o-mini	0.0	0.9	128,000
Claude 3.5 Sonnet	0.0	0.9	4,096
Claude 4.5 Sonnet	0.0	0.9	200,000
Claude 4.5 Haiku	0.0	0.9	200,000
Qwen3.5-9B	0.0	0.9	262,144

Table 9: 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 10: 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.

J Ablation Study on Verifier

Table 14 presents the evaluation results of refined and non-refined questions comparison in the verification process.

K Ablation Study on Complete Multi-hop

Table 15 presents the complete ablation study results on multi-hop questions.

Table 16 and 17 present the results of lexical overlap analysis.

Table 18 presents the results of cross-model stress test on GPT-4o and GPT-4o-mini.

L Error Category Details

L.1 Error Category Definitions

There are two primary error category and each of them has two subcategories. Each subcategory has three criteria to be classified as its subcategory. Table 19 demonstrates the definition of each subcategory.

L.2 Error Category Examples

Figure 4 demonstrates the example of each subcategory of RAG failure mode.

L.3 Evaluation of Error Category

The table 20 demonstrates the inter-annotator agreement and performance of different comparison methods for evaluating the error category of RAG failure mode.

L.4 Error Category Distribution

Table 21 demonstrates the distribution of error categories in RAG systems across different models, prompting strategies, and retriever configurations. The table provides a comparative analysis of the prevalence of each error category under standard

Retriever	Chunk	Gemma2-9B		Llama3.1-8B		Ministral-8B		Qwen3.5-9B		
		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 11: 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 12: 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	
Dataset	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
Sparse	79.9	63.1	71.2	59.2	74.2	59.0	82.7	62.3
Dense	84.1	63.6	75.1	58.8	77.8	57.9	85.0	62.4
Oracle	89.4	69.7	79.5	64.8	85.5	66.0	88.0	70.4

Table 13: RAG performance comparison between VERGE and Guinet dataset across four models (Gemma2-9B, Llama3.1-8B, Ministral-8B, Qwen3.5-9B) and four retrieval configurations (Closed-book, Sparse, Dense, Oracle). Values represent accuracy in %.

Knowledge Boundary Errors

Context Utilisation:
Q: Considering the challenges faced by both the NFL and BC athletics, what strategic solution, drawing from the NFL's approach, could potentially address BC's attendance issues while simultaneously enhancing the fan experience?
Answer Candidates:
A) Implementing a stricter blackout policy to incentivize full stadium attendance
B) Replacing bleachers with individual seats to increase capacity and improve comfort
C) Offering discounted season tickets to attract a wider range of fans
D) Partnering with local businesses to create exclusive pre-game tailgating experiences
Model answer: "C"
Correct answer: "B"
Exam taker reasoning: The NFL has faced attendance issues in the past, but they have addressed them through various strategies. One effective strategy is offering discounted season tickets to attract a wider range of fans. This can help increase attendance and revenue.
Source document: The NFL is lowering the bar, so that blackout rules don't require sellouts. Blackouts are not an issue in college, but perhaps few seats will help demand and make Alumni seem full. I don't want to tear out seats, but maybe we can replace the bleachers with actual seats. That would take up more space, eliminate seats and improve the watching experience.
Explanation: The exam taker incorrectly assumes that the NFL's solution to attendance issues is offering discounted season tickets. The provided text mentions lowering the blackout threshold and replacing bleachers with individual seats, not discounted tickets.

Knowledge Scope:
Q: What is the primary factor that the Patent Trial and Appeal Board (PTAB) considers when determining the patent eligibility of a method claim under Section 101?
Answer Candidates:
A) The presence of a tangible output device
B) The method's application to a specific industry
C) The method's ability to transform physical subject matter
D) The method's novelty and non-obviousness
Model answer: "A"
Correct answer: "C"
Exam taker reasoning: This is based on the Alice Corp. v. CLS Bank International decision, which established that a method claim must be more than just an abstract idea or law of nature to be patentable.
Source document: The claims should affirmatively claim the device, machine or component performing each step or function.
Explanation: The exam taker incorrectly attributes the 'transforming physical subject matter' requirement to the *Alice* decision. While *Alice* is relevant to Section 101, it focuses on distinguishing abstract ideas from patent-eligible inventions, not specifically on the transformation requirement.

Information Processing Errors

Reasoning:
Q: Given the tradeoffs presented in the figures for various elliptic curve and RSA cryptographic schemes, what is the minimum number of physical qubits required to break a cryptographic scheme with a classical security parameter of 192 bits within a 24-hour timeframe, assuming a physical error rate per gate of 10^{-3}?
Answer Candidates:
A) $1.28\\times 10^7$
B) $2.27\\times 10^8$
C) $7.70\\times 10^{10}$
D) $3.37\\times 10^7$
Model answer: "A"
Correct answer: "B"
Exam taker reasoning: To break a cryptographic scheme with a classical security parameter of 192 bits, we need to consider the number of qubits required for Shor's algorithm. The number of qubits required is given by the formula $2^{n/2}$, where n is the security parameter. For a 192-bit security parameter, the number of qubits required is $2^{192/2} = 2^{96}$. However, we also need to consider the physical error rate per gate of 10^{-3}. The number of qubits required to achieve a 24-hour timeframe with this error rate is $2^{96} \times 10^{-3} = 2^{96.5}$. The closest answer to this value is 2.27×10^8.
Source document: The classical security parameter is 192 bits.
Explanation: The exam taker incorrectly applies a formula for the number of qubits required for Shor's algorithm, assuming a direct relationship between security parameter and qubit count. They also introduce an arbitrary scaling factor based on the error rate without justification from the provided context.

Information Integration:
Q: Based on the provided drilling data, what can be inferred about the geological structure and potential for mineral exploration in the area west of the known mineralization?
Answer Candidates:
A) The presence of altered volcanic rock and a fault zone suggests a high likelihood of finding significant precious metal deposits in this area
B) The absence of significant precious metal values in hole 05-07 indicates that the West Cliffs vein is unlikely to be present in this region
C) The thicker alluvial cover encountered in hole 05-05 may have obscured the presence of the West Cliffs vein, warranting further exploration in this area
D) The presence of an opaline rich structure with sulfide minerals and low temperature quartz in hole 05-07 suggests a potential for hydrothermal alteration and mineralization in this area
Model answer: "A"
Correct answer: "D"
Exam taker reasoning: he drilling data suggests that the area west of the known mineralization has altered volcanic rock and a fault zone, which could indicate a high likelihood of finding significant precious metal deposits.
Source document: The hole intersected an opaline rich structure hosted within basalt and andesite west of any previous drilling. The structure contained sulfide minerals (mainly pyrite) and low temperature quartz, but did not contain any significant precious metal values.
Explanation: The exam taker correctly identifies altered volcanic rock and a fault zone as potential indicators of mineralization. However, they fail to integrate the information from hole 05-07, which indicates the presence of these features without significant precious metal values.

Figure 4: Example of each subcategory of RAG failure mode.

Metric	Refined	Non-Refined
MH validity	94.67%	95.00%
QA integrity	96.67%	97.00%
Distractor quality	3.55	3.22

Table 14: Evaluation of Refined vs. Non-Refined Questions

and Chain-of-Thought (CoT) prompting conditions, with a chunk size of 4,000 and a sparse retriever. The distribution highlights the impact of different configurations on the types of errors made by RAG systems, offering insights into areas for improvement in model performance.

M Open-ended Question Answer Generation

M.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) → Understanding(B) → Explanation(C)
2. **Convergent Reasoning**: Multiple chunks provide complementary aspects converging to unified conclusion
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

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

Table 15: 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 16: 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 17: Accuracy by Lexical Overlap Quartile

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

M.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

4. Improvement Recommendations

- Multi-hop failure: Suggest comprehensive chunk incorporation strategies
- Answer incompleteness: Recommend completeness-enhancing details

Primary	Subcategory	Definition
Information Processing	Integration	- Incorrect combination of retrieved facts - Failure to establish relationships between facts - Missing key connections between pieces of information
	Reasoning	- Invalid logical inference steps - Incorrect causality assumptions - Faulty deductive or inductive reasoning
Knowledge Boundary	Context utilisation	- Missing or overlooking relevant information - Using irrelevant or misleading information - Failure to identify key supporting evidence
	Knowledge scope	- Making assumptions beyond the given information - Introducing external knowledge inappropriately - Misinterpreting information boundaries

Table 19: Definitions of each error category

Comparison	kappa	Accuracy
Human (Fleiss' κ)	0.77	89%
Human (majority vote)	0.97	99%
Sonnet 4.5	0.80	94%

Table 20: Inter-annotator Agreement and Performance of Different Comparison Methods

- 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

M.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

		Gemma2-9B				Llama3.1-8B				Ministral-8B			
		Standard		CoT		Standard		CoT		Standard		CoT	
Primary	**Subcategory**	#	%	#	%	#	%	#	%	#	%	#	%
Knowledge	Context Utilisation	359	53.8	425	76.2	376	53.0	287	63.5	409	54.4	377	58.6
Boundary	Knowledge Scope	174	26.1	95	17.0	182	25.6	101	22.3	189	25.1	168	26.1
Information	Integration	94	14.1	16	2.9	107	15.1	34	7.5	113	15.0	57	8.9
Processing	Reasoning	40	6.0	22	3.9	45	6.3	30	6.6	41	5.5	41	6.4
Total Categorised Errors		667		558		710		452		752		643	

Table 21: Distribution of Error Categories in RAG Systems: Comparative Analysis of Standard versus Chain-of-Thought (CoT) Prompting with chunk size of 4,000 and sparse retriever.

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

M.4 Open-ended Question Answer Example

M.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.

M.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.

M.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.

M.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.

M.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.