

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

Anonymous ACL submission

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 10,243 candidate questions spanning five domains, VERGE filters or refines low-quality items, yielding 2,258 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. Our methodology provides both a practical RAG evaluation suite for practitioners and a rigorous foundation for advancing multi-hop reasoning research.

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.

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. Across 10,243 candidate questions spanning five domains, VERGE filters or refines low-quality items, yielding 2,258 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 ($p < 0.001$).

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 (integration failures and reasoning failures) and Knowledge Boundary Errors (context utilisation failures and knowledge scope violations). 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 sys-

083	tems, validated through comprehensive human assessment and experiments across five diverse task domains.	133
084		134
085		135
086	• We propose a hierarchical error taxonomy specifically designed for multi-hop RAG system analysis.	136
087		137
088		138
089	2 Related Literature	139
090	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.	140
091		141
092		142
093		143
094		144
095		145
096		146
097		147
098		148
099		149
100		150
101		151
102		152
103		153
104		154
105		155
106		156
107		157
108		158
109		159
110		160
111		161
112		162
113		163
114		164
115		165
116		166
117		167
118		168
119		169
120		170
121		171
122		172
123		173
124		174
125		175
126		176
127		177
128		178
129		179
130		180
131		181
132		182

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 $D = \{d_1, d_2, \dots, d_n\}$ represent a corpus of documents where each document d_i contains potentially relevant information. We chunk D into semantic chunks $C = \{c_1, c_2, \dots, c_m\}$, where each chunk c_j is a semantically coherent chunk with a maximum size of 4,000 characters. A multi-hop ques-

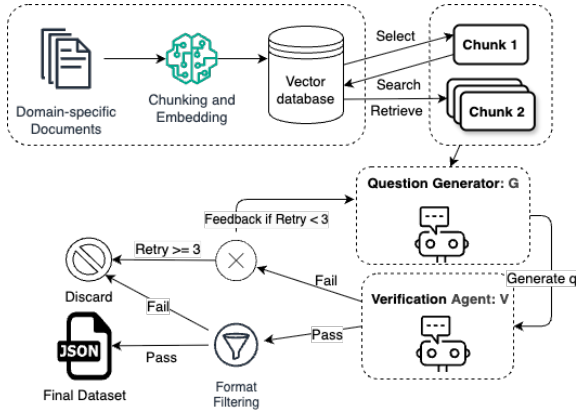


Figure 1: VERGE dataset generation process.

tion 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:

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

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 the multiple-choice item-writing taxonomy guidelines, which are derived from both theoretical and empirical research Haladyna et al. (2002) on creating effective distractors, our prompt instructs models to generate three types of distractors: *partial information distractors* that leverage incomplete chunk information understanding, *contextual misalignment distractors* that exploit potential misinterpretation patterns, and *critical dependency distractors* that appear correct if crucial inter-chunk dependencies are overlooked. Our approach demonstrates that the well-constructed prompt en-

Methodology	Timing	Feedback	Criteria
Es et al. (2024)	None	None	None
Guinet et al. (2024)	Post	None	IRT model
VERGE (Ours)	During	≤ 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.

gineering 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.

Domain	# Question	# Chunk	Size	Voc. Count
GovReport	614	5,152	326	24,963
HotpotQA	309	5,433	347	74,440
MultiFieldQA	559	1,640	344	31,973
2Wiki	467	2,884	347	49,145
SEC Filings	309	36,834	337	52,968
Total	2,258	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 3.2-2B (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: Llama3.2-2B, Gemma2-9B, and Ministral-8B of varying sizes and architectures.

Our dataset consists of 2,258 multi-hop questions across five different task domains: $T = \{t_{\text{gov}}, t_{\text{hot}}, t_{\text{multi}}, t_{\text{wiki}}, t_{\text{sec}}\}$. 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 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 1,640 document chunks in t_{multi} to 36,834 in t_{sec} , with an average chunk size of 338 words. Our iterative verification framework yielded an acceptance rate of 22%, resulting in 2,258 verified questions from 10,243 candidates, with 267 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 three open language models; Llama3.1-8B (Dubey et al., 2024), Ministral-8B, and Gemma2-9B 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. To investigate the relationship between chunk size and RAG performance, we explore three chunk size configurations: 512, 2,048, and 4,000 characters with top k = 5.

Furthermore, this experiment explores whether an advanced prompting strategy designed for multi-hop reasoning tasks can enhance RAG system performance. We evaluate an established prompting technique: Chain-of-Thought (CoT) (Wei et al., 2022). This technique has demonstrated effectiveness on multi-hop reasoning tasks in previous research (Wang et al., 2023; Zhuang et al., 2024). Including this prompting strategy serves two purposes: it provides valuable baselines for contextualising our dataset results, and it helps us understand how RAG systems can be optimised for multi-hop queries through prompt engineering. 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.

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 to understand the granular classification of reasoning failures. The taxonomy comprises two primary categories: **Information Processing Errors** and **Knowledge Boundary Errors**. Information processing errors encom-

	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.

pass **integration failures**, which fail to establish relationships between facts, and **reasoning failures**, which are invalid logical inference steps. Knowledge boundary errors include **context utilisation** errors, which use irrelevant or misleading information and **knowledge scope violations**, which are inappropriate introductions of external knowledge.

We employ Gemma2-9B as our error classifier in this experiment. It takes a QA pair and the reasoning of the selected answer to classify it into a primary error category and a subcategory along with the explanation of the classification.

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.

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).

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) ($\chi^2 = 12.8$, $p < 0.001$). Additionally, VERGE exhibits better overall distractor quality, averaging 3.219 (± 0.683), compared to 2.526 (± 1.047) for the Guinet dataset, as assessed via the Mann-Whitney U test (Mann and Whitney, 1947) ($U = 64,045.5$, $p < 0.001$).

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. Firstly, almost all RAG configurations except Llama3.1-8B with CoT prompting 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. Secondly, 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 63.8%) with the standard prompting compared to CoT prompting (up to 61.6%) for chunk size of up to 2,048, yet CoT prompting outperforms standard prompting (64.1% vs 63.8%)

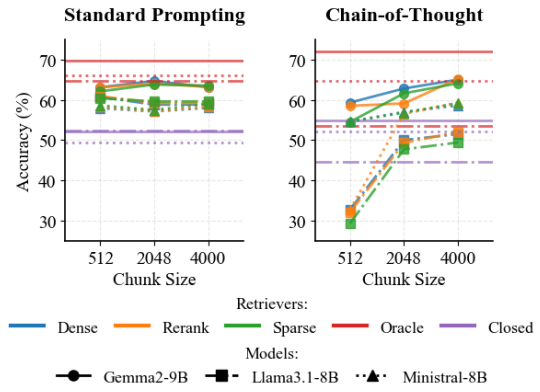


Figure 2: Accuracy in %. Results across three LLMs, three chunk sizes (512, 2,048, 4,000 characters), three retrievers, and two prompting strategies (Standard and Chain-of-Thought), plus closed-book and Oracle settings.

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. However, CoT prompting introduces parsing failures, where the generated response is not in the expected format, especially for Llama3.1-8B, which exhibits the highest failure rates of up to 50% at the worst case. The full results, including accuracies excluding the parsing failures, are reported in Appendix I.

4.3 Error Category Analysis

Table 5 demonstrates the error category distribution for RAG systems across standard and CoT prompting with sparse retriever and the chunk size of 4,000 characters. 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 consistent 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.

When comparing prompting strategies, CoT

498	prompting reduces information processing errors,	fully refined through verification feedback. Out of	548
499	particularly integration failures, while the propor-	267 refined questions, 201 questions were refined	549
500	tion of reasoning errors remains relatively stable	with one iteration and 66 questions required two	550
501	for Llama3.1-8B and Ministral-8B models. Con-	iterations.	551
502	versely, context utilisation errors increase with CoT		
503	prompting, especially for Gemma2-9B - rising	5.2 Ablation of Complete Multi-Hop	552
504	from 53.8% to 76.2%. This suggests that while	The complete multi-hop criterion is evaluated by	553
505	CoT enhances reasoning structure, it may lead mod-	ablating chunks from the context used to answer	554
506	els to become more selective about incorporated in-	each question in the VERGE dataset. We observe	555
507	formation, potentially overlooking crucial evidence.	accuracy degradation of 11.8% for the single best	556
508	These patterns explain why even in Oracle settings,	chunk was fed and 4.8% for a single chunk was	557
509	models struggle to effectively synthesise informa-	removed from the full context on average com-	558
510	tion across multiple sources, highlighting context	pared to the oracle setting. Partial subsets yield	559
511	utilisation as the primary challenge in multi-hop	modest improvement over the closed-book base-	560
512	RAG systems.	line, which indicates limited local shortcutability.	561
		However, the pronounced performance gap relative	562
513	5 Analysis and Discussion	to the oracle setting indicates that no subset alone	563
		is sufficient to reliably resolve the question. These	564
514	5.1 Ablation of Verifier	findings highlight that multi-hop completeness in	565
		VERGE reflects distributed evidence dependence	566
515	The verification agent relies on two criteria to filter	rather than strict per-chunk indispensability.	567
516	and refine the dataset: complete multi-hop and QA		
517	integrity. To assess the effectiveness of individual	5.3 Cost-Efficiency Analysis	568
518	verification criteria, we observe the number of elim-	Table 7 reports the average wall time per iteration	569
519	inated questions per verification criteria in Table	depth and monetary cost. As practitioners often	570
520	6. The complete multi-hop criteria demonstrates	expected to use API-accessible frontier models	571
521	significantly higher filtering impact, eliminating	rather than local LLMs, we extend our analysis to	572
522	74.2% of candidates from the initial dataset. This	Claude 4.5 Sonnet (Anthropic, 2025b) and Haiku	573
523	reflects the difficulty of generating questions that re-	(Anthropic, 2025a). Across all models, generation	574
524	quire information synthesis across multiple chunks,	completes within 6 - 45 seconds depending on the	575
525	even with iterative refinement opportunities. In	model size and number of iterations. API usage	576
526	contrast, the QA integrity criteria filters a compar-	is similarly modest, costing under \$5 per 100 QA	577
527	atively modest 8.2% of questions, suggesting that	pairs for Sonnet and under \$2 for Haiku. Overall,	578
528	generating a question with factual consistency is	VERGE remains practical and cost-efficient in real-	579
529	less challenging to maintain than multi-hop reason-	world workflows, particularly when paired with	580
530	ing requirements. Despite their asymmetric filter-	stronger generator models.	581
531	ing effects, both criteria demonstrate complemen-		
532	tary value in maintaining dataset quality. While	5.4 Open-Ended Question Generation	582
533	significant overlap exists in their filtering decisions,	While our main evaluation focuses on multiple-	583
534	each constraint captures unique quality dimensions.	choice questions, VERGE framework naturally ex-	584
535	Figure 3 illustrates the refinement process, show-	tends to open-ended questions. The open-ended	585
536	ing how the feedback mechanism guides the	variant maintains our verification mechanism, en-	586
537	incorporation of critical information from an un-	suring multi-hop reasoning requirements while gen-	587
538	used chunk, transforming what would otherwise be	erating comprehensive reference answers. Out	588
539	rejected candidates into a valid multi-hop question.	of 600 generated open-ended question candidates,	589
540	These analyses exhibit both filter and refine capabil-	56.0% of QA pairs passed the verification, which	590
541	ities of the verification agent represent an integral	is notably higher than the 23.1% acceptance rate	591
542	part of VERGE framework and differentiate our	for multiple-choice questions. This difference re-	592
543	method to other existing RAG evaluation dataset	fects the inherent complexity of distractor genera-	593
544	generation methods.	tion rather than multi-hop reasoning requirements,	594
545	In terms of the effectiveness of the refinement	highlighting that distractor creation remains a fun-	595
546	process, our analysis reveals that 267 questions,	damental challenge. The prompt and examples of	596
547	which is 11.8% of the final dataset were success-		

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

Table 5: 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.

<p>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</p>	<p>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</p>
--	---

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

Criteria	# eliminated	% eliminated
Complete Multi-Hop	7,598	74.2%
QA Integrity	841	8.2%
Both Criteria	7,758	75.7%
Format Filtering	227	2.2%

Table 6: Number of eliminated questions per verification criteria and final formatting filtering. Initially there are 10,243 candidate questions and survived 2,258 questions.

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 7: 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.

open-ended questions generated by VERGE are found in the appendix L.

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.

By offering a principled mechanism for generating high-quality, domain-specific evaluation data and a taxonomy that sharpens our understanding of RAG failure modes, VERGE provides a practical tool for applied NLP and a foundation for advancing the science of multi-hop reasoning evaluation.

7 Limitations

There are important inherent limitations to consider in this study. First, as both our method and experiments rely on LLMs, the quality of the dataset and experimental outcomes is directly influenced by the underlying model capabilities. Due to resource constraints, our experiments employ relatively small open models due to the resource limitation. This choice may lead to outcomes that differ from those obtained with larger state-of-the-art (SOTA) models, which typically exhibit substantially stronger generation and refinement abilities. Specifically, larger SOTA models would likely demonstrate higher-quality dataset generation and improved survival rates of generated questions due to enhanced generation capabilities and more effective refinement. When generating 100 QA pairs, the Claude 4.5 models (Sonnet and Haiku) required no refinement for 55% and 48% of questions, substantially higher than the 29% observed with local open-weight models. They also required far fewer two-iteration refinements, indicating higher initial question quality and reduced refinement overhead.

However, our human annotation study shows that the VERGE-generated dataset demonstrates higher-quality QA pairs compared to the Guinet dataset, which was generated using a larger model. Thus, this suggests that the VERGE method would remain effective and potentially generate even higher quality datasets when paired with larger SOTA models. Additionally, while the distribution of error categories might differ, we hypothesise that the primary failure patterns identified would persist across model scales, albeit at different proportions.

Second, VERGE enforces an operational notion of Complete Multi-Hop in which correct answering depends meaningfully on integrating multiple evidence chunks: while partial subsets may provide limited assistance, no strict subset achieves performance comparable to the full-context condition, and chunk removal leads to systematic accuracy degradation.

Third, while our experiments with CoT prompting demonstrated RAG performance improvements for specific model configurations, we did not explore these prompting techniques in the dataset generation framework. Given the observed effectiveness of such prompting strategies, integrating them into the dataset generation and verification process could potentially lead to higher-quality questions. This prompting technique let LLMs logically anal-

yse the relationships of multi-document information across the chunks.

Fourth, evaluating dataset quality remains a fundamental methodological challenge. We define dataset quality in terms of discriminative power, yet this introduces potential circularity, as the dataset's value is inferred from the same model performance it is used to assess. This limitation parallels the IRT-based approach (Guinet et al., 2024), which optimises for informativeness post hoc via the IRT model rather than relying on independent quality metrics.

Fifth, our error category taxonomy is derived from observations of our RAG experiments. As RAG architectures continue to evolve, new error categories or subcategories may emerge that our current framework does not capture.

Finally, although this work focuses on multiple choice question generation, VERGE method can easily be extended to open-ended question generation as the analysis has shown in the section 5.4.

These limitations suggest several future work directions. Leveraging larger SOTA models could improve both the dataset generation process and the effectiveness of the verification agent. In tandem, integrating advanced prompting strategies into both dataset construction and verification may further improve dataset quality. It would likely result in more complex and higher-quality multi-hop questions. Finally, generalising the error taxonomy to accommodate evolving failure patterns in RAG systems would enhance its analytical robustness and adaptability.

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.

722	Jacob Cohen. 1960. A coefficient of agreement for nominal scales. <i>Educational and psychological measurement</i> , 20(1):37–46.	
723		
724		
725	Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The Llama 3 Herd of Models . Preprint, arXiv:2407.21783.	
726		
727		
728	Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation . In <i>Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations</i> , pages 150–158, St. Julians, Malta. Association for Computational Linguistics.	
729		
730		
731		
732		
733		
734		
735	Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. <i>Psychological bulletin</i> , 76(5):378.	
736		
737		
738	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. <i>arXiv preprint arXiv:2502.20377</i> .	
739		
740		
741		
742		
743		
744	Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated evaluation of retrieval-augmented language models with task-specific exam generation . In <i>Proceedings of the 41st International Conference on Machine Learning</i> , volume 235 of <i>Proceedings of Machine Learning Research</i> , pages 16773–16801. PMLR.	
745		
746		
747		
748		
749		
750		
751	Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. <i>Applied measurement in education</i> , 15(3):309–333.	
752		
753		
754		
755		
756	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 <i>Proceedings of the 28th International Conference on Computational Linguistics</i> , pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.	
757		
758		
759		
760		
761		
762		
763	Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 1419–1436, Online. Association for Computational Linguistics.	
764		
765		
766		
767		
768		
769		
770	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 <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)</i> , pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.	
771		
772		
773		
774		
775		
776		
777		
778		
	Jeongsoo Lee, Daeyong Kwon, and Kyohoon Jin. 2025a. Grade: Generating multi-hop qa and fine-grained difficulty matrix for rag evaluation . In <i>Findings of the Association for Computational Linguistics: EMNLP 2025</i> , pages 4405–4424.	779
		780
		781
		782
		783
	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. <i>arXiv preprint arXiv:2504.08756</i> .	784
		785
		786
		787
		788
	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. <i>Advances in neural information processing systems</i> , 33:9459–9474.	789
		790
		791
		792
		793
		794
		795
	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 <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track</i> , pages 881–893, Miami, Florida, US. Association for Computational Linguistics.	796
		797
		798
		799
		800
		801
		802
		803
	Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of tokens make the world go round . In <i>Proceedings of the Third Workshop on Economics and Natural Language Processing</i> , pages 13–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.	804
		805
		806
		807
		808
		809
		810
	Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. <i>The annals of mathematical statistics</i> , pages 50–60.	811
		812
		813
		814
	Meta. 2024. Model cards & prompt formats llama 3.2 . Accessed: 2026-01-02.	815
		816
	Mistral AI. 2024. Introducing Ministraux: The world’s best edge models . Accessed: 2026-01-02.	817
		818
	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 <i>Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 5336–5358, Seattle, United States. Association for Computational Linguistics.	819
		820
		821
		822
		823
		824
		825
		826
		827
		828
	Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations . In <i>Advances in Neural Information Processing Systems</i> , volume 37, pages 68772–68802. Curran Associates, Inc.	829
		830
		831
		832
		833

834	Karl Pearson. 1900. X. on the criterion that a given	Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-	890
835	system of deviations from the probable in the case	prompted chain-of-thought on large language mod-	891
836	of a correlated system of variables is such that it	els for open-domain multi-hop reasoning . In <i>Find-</i>	892
837	can be reasonably supposed to have arisen from	<i>ings of the Association for Computational Linguis-</i>	893
838	random sampling. <i>The London, Edinburgh, and</i>	<i>tics: EMNLP 2023</i> , pages 2717–2731, Singapore.	894
839	<i>Dublin Philosophical Magazine and Journal of Sci-</i>	Association for Computational Linguistics.	895
840	<i>ence</i> , 50(302):157–175.		
841	Nils Reimers and Iryna Gurevych. 2019. Sentence-	Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei	896
842	BERT: Sentence embeddings using Siamese BERT-	Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong,	897
843	networks . In <i>Proceedings of the 2019 Conference on</i>	Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large lan-	898
844	<i>Empirical Methods in Natural Language Processing</i>	guage models are not fair evaluators . In <i>Proceedings</i>	899
845	<i>and the 9th International Joint Conference on Natu-</i>	<i>of the 62nd Annual Meeting of the Association for</i>	900
846	<i>ral Language Processing (EMNLP-IJCNLP)</i> , pages	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	901
847	3982–3992, Hong Kong, China. Association for Com-	pages 9440–9450, Bangkok, Thailand. Association	902
848	putational Linguistics.	for Computational Linguistics.	903
849	Stephen Robertson and Hugo Zaragoza. 2009. The	Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan	904
850	probabilistic relevance framework: Bm25 and be-	Yang, and Ming Zhou. 2020. Minilm: Deep self-	905
851	yond . <i>Found. Trends Inf. Retr.</i> , 3(4):333–389.	attention distillation for task-agnostic compression	906
852	Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie	of pre-trained transformers . <i>ArXiv</i> , abs/2002.10957.	907
853	Huang, Nan Duan, and Weizhu Chen. 2023. En-	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	908
854	hancing retrieval-augmented large language models	Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,	909
855	with iterative retrieval-generation synergy . In <i>Find-</i>	and Denny Zhou. 2022. Chain-of-thought prompt-	910
856	<i>ings of the Association for Computational Linguis-</i>	ing elicits reasoning in large language models . In	911
857	<i>tics: EMNLP 2023</i> , pages 9248–9274, Singapore.	<i>Proceedings of the 36th International Conference on</i>	912
858	Association for Computational Linguistics.	<i>Neural Information Processing Systems, NIPS ’22</i> ,	913
859	Zhiyu Shen, Jiyuan Liu, Yunhe Pang, and Yanghui Rao.	Red Hook, NY, USA. Curran Associates Inc.	914
860	2025. Hopweaver: Synthesizing authentic multi-	Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas	915
861	hop questions across text corpora. <i>arXiv preprint</i>	Muennighoff. 2023. C-pack: Packaged resources	916
862	<i>arXiv:2505.15087</i> .	to advance general chinese embedding . <i>Preprint</i> ,	917
863	Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao,	<i>arXiv:2309.07597</i> .	918
864	Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024.	Yunqi Xu, Tianchi Cai, Jiyan Jiang, and Xierui Song.	919
865	Generate-then-ground in retrieval-augmented genera-	2024. Face4rag: Factual consistency evaluation for	920
866	tion for multi-hop question answering . In <i>Proceed-</i>	retrieval augmented generation in chinese . In <i>Pro-</i>	921
867	<i>ceedings of the 62nd Annual Meeting of the Association</i>	<i>ceedings of the 30th ACM SIGKDD Conference on</i>	922
868	<i>for Computational Linguistics (Volume 1: Long Pa-</i>	<i>Knowledge Discovery and Data Mining, KDD ’24</i> ,	923
869	<i>pers)</i> , pages 7339–7353, Bangkok, Thailand. Associ-	pages 6083–6094, New York, NY, USA. Association	924
870	ation for Computational Linguistics.	for Computing Machinery.	925
871	Sebastian Simon, Alina Mailach, Johannes Dorn, and	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio,	926
872	Norbert Siegmund. 2024. A methodology for evaluat-	William Cohen, Ruslan Salakhutdinov, and Christo-	927
873	ing rag systems: A case study on configuration depen-	pher D. Manning. 2018. HotpotQA: A dataset for	928
874	dency validation. <i>arXiv preprint arXiv:2410.08801</i> .	diverse, explainable multi-hop question answering .	929
875	Yixuan Tang and Yi Yang. 2024. Multihop-RAG:	In <i>Proceedings of the 2018 Conference on Empiri-</i>	930
876	Benchmarking retrieval-augmented generation for	<i>cal Methods in Natural Language Processing</i> , pages	931
877	multi-hop queries . In <i>First Conference on Language</i>	2369–2380, Brussels, Belgium. Association for Com-	932
878	<i>Modeling</i> .	putational Linguistics.	933
879	Gemma Team, Morgane Riviere, Shreya Pathak,	Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao,	934
880	Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupati-	Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and	935
881	raju, Léonard Hussenot, Thomas Mesnard, Bobak	Lei Li. 2025. Rare: Retrieval-aware robustness eval-	936
882	Shahriari, Alexandre Ramé, and 1 others. 2024.	uation for retrieval-augmented generation systems .	937
883	Gemma 2: Improving open language models at a	<i>arXiv preprint arXiv:2506.00789</i> .	938
884	practical size. <i>arXiv preprint arXiv:2408.00118</i> .	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	939
885	Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot,	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	940
886	and Ashish Sabharwal. 2022. MuSiQue: Multi-	Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang,	941
887	hop questions via single-hop question composition .	Joseph E Gonzalez, and Ion Stoica. 2023. Judging	942
888	<i>Transactions of the Association for Computational</i>	llm-as-a-judge with mt-bench and chatbot arena . In	943
889	<i>Linguistics</i> , 10:539–554.	<i>Advances in Neural Information Processing Systems</i> ,	944
		volume 36, pages 46595–46623. Curran Associates,	945
		Inc.	946

947	Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024. Metacognitive retrieval-augmented large language models. In <i>Proceedings of the ACM Web Conference 2024</i> , pages 1453–1463.	
948		
949		
950		
951	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. <i>RAGEval: Scenario specific RAG evaluation dataset generation framework</i> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.	
952		
953		
954		
955		
956		
957		
958		
959		
960	Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. <i>EfficientRAG: Efficient retriever for multi-hop question answering</i> . In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 3392–3411, Miami, Florida, USA. Association for Computational Linguistics.	
961		
962		
963		
964		
965		
966		
967		
968		
	A Risk and Ethical Considerations	
969	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.	
970		
971		
972		
973		
974		
975		
976		
977		
978		
979		
980		
981		
982		
983		
984		
985		
986		
987		
	A.1 Licensing	
988	The source corpora are derived from LongBench (MIT License; Bai et al. (2023)) 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) and Claude models (Consumer Terms of Service) — permit research use.	
989		
990		
991		
992		
993		
994		
995		
996		
997		
	B Human Annotation Study Details	998
	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.	999
		1000
		1001
		1002
		1003
		1004
		1005
		1006
		1007
		1008
		1009
		1010
		1011
		1012
		1013
	B.1 Annotation Instruction	1014
	Task Overview	1015
	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:	1016
		1017
		1018
		1019
		1020
	• a question,	1021
		1022
	• four answer options (one correct answer and three distractors),	1023
		1024
	• a set of supporting document chunks.	1024
	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.	1025
		1026
		1027
		1028
		1029
		1030
		1031
	1. Verification Criteria	1032
	For each QA item, you will assess two binary criteria.	1033
		1034
	(a) Complete Multi-Hop Reasoning	1035
	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:	1036
		1037
		1038
		1039
	• The correct answer cannot be reliably determined using information from a single chunk alone.	1040
		1041
		1042
	• Multiple chunks must be jointly considered to resolve the question.	1043
		1044

1045	A question does not satisfy this criterion if:	• 4 — Very Good	1086
1046	• One chunk alone is reliably sufficient to answer the question correctly.	• 5 — Excellent (highly plausible and challenging)	1087
1047			1088
1048	• The question can be answered using general knowledge without integrating the provided chunks.	Please assign one the score per item.	1089
1049			1090
1050		3. Annotation Guidelines	1091
1051	Please answer Yes or No.	• Work independently and do not discuss your evaluations with others.	1092
1052			1093
1053	(b) Question-Answer (QA) Integrity	• If unsure, select the option that best reflects your judgment.	1094
1054	Determine whether the designated correct answer is fully supported by the provided document chunks. A question satisfies QA Integrity if:	• Please focus on accuracy and consistency.	1095
1055			1096
1056			1096
1057	• The correct answer is unambiguously derivable from the provided chunks.	All responses will be anonymised and used solely for research purposes.	1097
1058			1098
1059	• No factual inconsistencies or unsupported claims are present.	C Use of AI assistants	1099
1060		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.	1100
1061	A question does not satisfy this criterion if:		1101
1062	• The correct answer contradicts the provided context.		1102
1063			1103
1064	• Key information required to justify the correct answer is missing.		1104
1065			1105
1066	Please answer Yes or No.		1106
1067			1107
1068	2. Distractor Quality Assessment		1108
1069	For each QA item, evaluate the quality of the three incorrect answer options (distractors).		1109
1070	Rate the distractors on a 5-point Likert scale, considering:		1110
1071		D VERGE Prompts	1111
1072		D.1 Question Generation Prompt	1112
1073	• Relevance: How closely the distractors relate to the question and context.		1113
1074			1114
1075	• Plausibility: Whether the distractors appear reasonable given partial or incorrect reasoning.	Advanced Multi-Hop Question Generation Methodology Objective:	1115
1076		Generate an exceptionally challenging multiple-choice question (1 correct answer and 3 distractors) that demands sophisticated, interdependent reasoning across multiple document chunks.	1116
1077			1117
1078	• Difficulty: Whether the distractors would challenge a model or reader who has not fully integrated all required information.		1118
1079		Comprehensive Question Design Criteria:	1119
1080		1. Multi-Hop Reasoning Architecture	1120
1081	Scale definition:	• Construct questions that necessitate intricate information synthesis	1121
1082	• 1 — Very poor (irrelevant or obviously incorrect)	• Enforce mandatory cross-chunk logical dependencies	1122
1083		• Eliminate single-chunk solution pathways	1123
1084	• 2 — Poor		1124
1085	• 3 — Good		1125
			1126
			1127

1128	• Create a complex reasoning landscape that requires:	• Difficulty Level: Advanced (Graduate/Professional Examination Standard)	1175
1129			1176
1130	– Temporal reasoning		
1131	– Contextual cross-referencing		
1132	– Nuanced inference mechanisms		
1133	– Sophisticated information integration		
1134			
1135	2. Chunk Interdependency Strategies	5. Linguistic Precision Requirements	1177
1136	• Mandate that NO single document chunk reliably provides a complete answer	• Use sophisticated, academically rigorous language	1178
1137	• Implement strategic information distribution across chunks	• Avoid direct chunk referencing	1180
1138	• Create deliberate cognitive gaps requiring active reasoning	• Construct questions requiring active cognitive engagement	1181
1139	• Establish intricate logical bridges between disparate information sources	• Ensure questions are solvable exclusively through provided documentation	1182
1140			1183
1141			1184
1142		Output Format:	1185
1143		Question: [Sophisticated multi-hop reasoning challenge]	1186
1144	3. Distractor Engineering Principles	A) [Option A]	1187
1145	Three distractors must be meticulously crafted to exploit specific reasoning vulnerabilities:	B) [Option B]	1188
1146		C) [Option C]	1189
1147	a) Partial Information Distractor	D) [Option D]	1190
1148	• Leverages incomplete chunk understanding	Correct Answer: [Letter one of "A", "B", "C" or "D"]	1191
1149	• Appears superficially plausible	* <i>Do not generate anything else other than the above output format</i>	1192
1150	• Represents a common cognitive shortcut		1193
1151		Fundamental Constraints:	1194
1152	b) Contextual Misalignment Distractor	• Zero reliable single-chunk solution possibilities	1195
1153	• Based on seemingly logical but fundamentally flawed reasoning	• Mandatory full documentation utilization	1196
1154	• Requires subtle comprehension of inter-chunk relationships	• Rigorous reasoning complexity	1197
1155	• Exploits potential misinterpretation patterns	• Linguistically sophisticated presentation	1198
1156			1199
1157	c) Critical Dependency Distractor	Evaluation Heuristic:	1200
1158	• Appears correct if crucial inter-chunk dependencies are overlooked	A high-quality multi-hop question successfully prevents:	1201
1159	• Demands highest-order reasoning to distinguish from correct answer	• Surface-level information scanning	1202
1160	• Represents a sophisticated reasoning challenge	• Premature answer selection	1203
1161		• Incomplete reasoning processes	1204
1162			1205
1163		Domain: {task_domain}	1206
1164		Documentation: {documentation}	1207
1165			1208
1166		Generate an advanced multi-hop reasoning question	1210
1167	4. Reasoning Complexity Calibration		1211
1168	• Questions should require:		
1169	– Minimum {num_chunks}-step reasoning process		
1170	– Explicit integration of ALL provided chunks		
1171	– Ability to distinguish nuanced information relationships		
1172			
1173			
1174			

1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252

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

1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294

1295	• Provide comprehensive analysis of multi-hop reasoning requirements	• Introducing external knowledge inappropriately	1335
1296			1336
1297	• Ensure rigorous evaluation of question-answer integrity	• Misinterpreting information boundaries	1337
1298			1338
1299	• Maintain focus on document chunk utilization assessment		
1300			
1301	Perform comprehensive multi-hop reasoning verification	Analysis Instructions:	1339
1302		1. Identify error chains by tracing the progression of mistakes	1340
			1341
		2. If none of the categories are the right category, return “other”	1342
			1343
		3. For each error identified, provide:	1344
		• Evidence from the exam taker reasoning	1345
		• Specific text from source documents	1346
		• Explanation of how the error led to the incorrect answer	1347
			1348
1303	D.3 Error Categorisation Prompt		
1304	Multi-Hop Question Answering Error Analysis Methodology Objective:	Output Format:	1349
1305	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.	Return strictly only JSON format as specified below and do not return other explanations.	1350
1306			1351
1307			
1308			
1309			
1310	Primary Error Categories:	JSON Schema:	1352
1311	1. Information Processing Errors	• error_chain: Dictionary containing:	1353
1312	A. Integration Failures	– primary_error: Dictionary with:	1354
1313	• Incorrect combination of retrieved facts	* category: String (Main category from hierarchy)	1355
1314	• Failure to establish relationships between facts	* subcategory: String (Specific error type either A or B)	1356
1315	• Missing key connections between pieces of information	* confidence: Integer (1-5 scale)	1357
1316		* evidence: Dictionary containing:	1358
1317		· exam_taker_reasoning: String (Relevant quote)	1359
1318		· source_document: String (Relevant quote)	1360
1319	B. Reasoning Failures	· explanation: String	1361
1320	• Invalid logical inference steps	– contributing_errors: Array of dictionaries (Same structure as primary_error)	1362
1321	• Incorrect causality assumptions		1363
1322	• Faulty deductive or inductive reasoning	• quantitative_indicators: Dictionary containing:	1364
1323		– information_overlap: Float (Semantic similarity with source)	1365
1324	2. Knowledge Boundary Errors	– reasoning_steps_identified: Integer	1366
1325	A. Context Utilization	– external_knowledge_ratio: Float (Proportion of unsupported claims)	1367
1326	• Missing or overlooking relevant information		1368
1327	• Using irrelevant or misleading information		1369
1328	• Failure to identify key supporting evidence		1370
1329			1371
1330	B. Knowledge Scope		1372
1331	• Making assumptions beyond given information		1373
1332			1374
1333			1375
1334			

1376	Input Variables:	• Identify the choice with strongest documentary support	1416
1377	Question: {question}		1417
1378	Options:		
1379	A) {option_a}		
1380	B) {option_b}		
1381	C) {option_c}		
1382	D) {option_d}		
1383	Correct Answer: {correct_answer}		
1384	Documents: {chunk_text}		
1385	Exam Taker Reasoning:		
1386	{exam_taker_reasoning}		
1387	Fundamental Constraints:		
1388	• Generate only JSON response following the specified schema		
1389			
1390	• Maintain rigorous hierarchical error classification		
1391			
1392	• Provide comprehensive evidence-based analysis		
1393			
1394	• Ensure traceability of error progression chains		
1395	Perform comprehensive multi-hop error analysis		
1396			
<hr/>			
1397	D.4 Standard Prompt		
<hr/>			
1398	Multiple Choice Exam Taking Methodology Objective:		
1399	You are an AI assistant taking a multiple choice exam. Your task is to analyze the provided materials and select the most appropriate answer.		
1400			
1401			
1402			
1403	Examination Process:		
1404	1. Comprehensive Document Analysis		
1405	• Read the question, provided choices and documents carefully		
1406			
1407	• Extract relevant information from all provided documentation		
1408			
1409	• Identify key concepts and relationships within the materials		
1410			
1411	2. Choice Evaluation Framework		
1412	• Analyze each choice systematically against the documentation		
1413			
1414	• Assess the logical consistency of each option		
1415			
		3. Answer Selection Protocol	1418
		• Select the most appropriate answer based on evidence	1419
			1420
		• Ensure selection aligns with provided documentation	1421
			1422
		• Apply rigorous analytical reasoning to final choice	1423
			1424
	Input Variables:		1425
	Question: {question.question}		1426
	Choices:		1427
	{formatted_choices}		1428
	Documents:		1429
	{question.documentation}		1430
	Response Format Requirements:		1431
	• You must respond with exactly one letter: A, B, C, or D		1432
			1433
	• Do not include any explanation, period, or additional text		1434
			1435
	• Just the letter of the correct answer		1436
	Examples of Valid Responses:		1437
	• A		1438
	• B		1439
	• C		1440
	• D		1441
	Fundamental Constraints:		1442
	• Generate only single letter response		1443
	• Base selection exclusively on provided documentation		1444
			1445
	• Maintain strict adherence to response format		1446
	• Ensure analytical rigor in choice evaluation		1447
	Your answer (one letter only):		1448

1449

D.5 Chain-of-Thought prompt

1450

Step-by-Step Reasoning Methodology Objective:

1451

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>.

1452

1453

1454

1455

1456

1457

Response Format Requirements:

1458

1459

- Your answer should be after <Answer> in JSON format

1460

1461

- Use key “thought” and “answer” with string values

1462

1463

- The “answer” must be one of “A”, “B”, “C”, or “D”

1464

- Do not output anything else

1465

Reasoning Process:

1466

1. Document Analysis

1467

1468

1469

1470

1471

- Systematically examine the supporting documents
- Extract relevant information for the given question
- Identify key facts and relationships

1472

2. Step-by-Step Thinking

1473

1474

1475

1476

1477

1478

- Break down the question into component parts
- Connect document information to question requirements
- Apply logical reasoning to derive conclusions

1479

3. Answer Selection

1480

1481

1482

1483

1484

- Evaluate each choice against your reasoning
- Select the most appropriate option
- Ensure selection is supported by documentary evidence

1485

Example Responses:

1486

Example 1:

1487

1488

<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?

1489

1490

<Answer>:

1491

{{“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”}}

1492

1493

1494

1495

1496

1497

1498

1499

1500

Example 2:

1501

<Question>: Between Chinua Achebe and Rachel Carson, who had more diverse jobs?

1502

<Answer>:

1503

{{“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”}}

1504

1505

1506

1507

1508

1509

1510

1511

Example 3:

1512

<Question>: Remember Me Ballin’ is a CD single by Indo G that features an American rapper born in what year?

1513

1514

1515

<Answer>:

1516

{{“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”}}

1517

1518

1519

1520

1521

Input Variables:

1522

Supporting Document:

1523

{question.documentation}

1524

Question:

1525

<Question>: {question.question}

1526

<Choices>: {formatted_choices}

1527

<Answer>:

1528

Fundamental Constraints:

1529

- Generate only JSON response with specified format

1530

1531

- Provide comprehensive step-by-step reasoning in “thought” field

1532

1533

- Select answer exclusively from provided choices (A, B, C, or D)

1534

1535

1536	• Base all reasoning on supporting documenta-	B) West Palm Beach-Boca Raton-Boynton Beach	1572
1537	tion	Metropolitan Statistical Area	1573
1538	Provide step-by-step reasoning and answer	C) Port Saint Lucie MSA	1574
1539	selection	D) Sebastian-Vero Beach MSA	1575
<hr/>			
1540	E Question Examples	Correct Answer: A	1576
1541	Below are arbitrary selected generated question per	E.2.1 Example MultiFieldQA Question	1577
1542	task domain in the dataset.	<hr/>	
1543	E.1 Example GovReport Question	Question:	1578
<hr/>			
1544	Question:	Given the Shadoks' optimization strategies and the	1579
1545	Considering the multifaceted approaches to ad-	BDFS algorithm's purpose, what is the most likely	1580
1546	dressing Puerto Rico's fiscal challenges outlined	reason the Shadoks observed a 30 solution improve-	1581
1547	in the provided documentation, what is the most	ment near the end of the challenge after incorporat-	1582
1548	significant factor hindering the successful imple-	ing BDFS?	1583
1549	mentation of long-term economic solutions?	Choices:	1584
1550	Choices:	A) BDFS effectively identified and eliminated	1585
1551	A) The lack of a comprehensive, integrated infor-	redundant color classes, leading to a more	1586
1552	mation technology system for financial man-	efficient coloring scheme.	1587
1553	agement.	B) BDFS's ability to recolor vertices at larger	1588
1554	B) The Oversight Board's reluctance to certify	distances from the conflict vertex allowed for	1589
1555	updated fiscal plans due to the impact of Hur-	a more nuanced exploration of the color space.	1590
1556	ricanes Irma and Maria.	C) BDFS's conflict minimization phase directly	1591
1557	C) The Puerto Rican government's dependence	addressed the limitations of the Shadoks' ini-	1592
1558	on federal assistance and its inability to gener-	tial conflict set management strategy.	1593
1559	ate sufficient revenue.	D) BDFS's use of a Gaussian random variable for	1594
1560	D) The absence of a clear consensus among fed-	the weight function (f) significantly improved	1595
1561	eral agencies regarding the most effective	the exploration of the solution space.	1596
1562	forms of technical assistance.	Correct Answer: B	1597
1563	Correct Answer: C	E.2.2 Example SecFilings Question	1598
<hr/>			
1564	E.2 Example HotpotQA Question	Question:	1599
<hr/>			
1565	Question:	Considering the evolution of U.S. assistance to	1600
1566	Which metropolitan area, based on commuting pat-	Georgia, what was the primary shift in focus ob-	1601
1567	terns, encompasses the largest population within	served between the period immediately follow-	1602
1568	the broader "Gold Coast" region of Florida?	ing the 2008 war and the years from FY2010 to	1603
1569	Choices:	FY2017?	1604
1570	A) Miami-Fort Lauderdale-Miami Beach	Choices:	1605
1571	Metropolitan Statistical Area	A) A transition from humanitarian aid to long-	1606
		term economic development initiatives.	1607

- 1608 B) A shift from military capacity building to pro-
1609 moting democratic reforms and civil society.
- 1610 C) A move from direct budgetary support to
1611 funding specific security and stabilization pro-
1612 grams.
- 1613 D) A change from supporting Georgia’s partici-
1614 pation in international peacekeeping missions
1615 to bolstering its border security.

1616 **Correct Answer: C**

1617 E.2.3 Example 2Wiki Question

1618 **Question:**

1619 What motivated Louisa Ulrika’s opposition to the
1620 marriage between Gustav and Sophia Magdalena?

1621 **Choices:**

- 1622 A) She believed Sophia Magdalena was unsuit-
1623 able due to her perceived lack of intelligence.
- 1624 B) She desired a political alliance with
1625 Brandenburg-Schwedt through a marriage
1626 between Gustav and her niece.
- 1627 C) She resented the Swedish public’s preference
1628 for Sophia Magdalena over her own preferred
1629 candidate.
- 1630 D) She felt the Swedish Queen exerted undue
1631 influence over the decision-making process.

1632 **Correct Answer: B**

1633 F Verification Process Survival Rate

1634 Table 8 presents the survival rates of generated
1635 question-answer pairs across the three LLMs used
1636 for generation.

1637 However, the number of surviving questions
1638 varies across task domains, with the lowest survival
1639 rate observed in the HotpotQA dataset (16.7%).
1640 This discrepancy is likely due to the inherent
1641 complexity of the HotpotQA dataset, which re-
1642 quires synthesising information across multiple
1643 documents to answer questions correctly. The Gov-
1644 Report dataset, on the other hand, has the highest
1645 survival rate (32.9%), indicating that the questions
1646 generated in this domain are more straightforward
1647 and less reliant on multi-hop reasoning.

	Gemma2-9B	Llama3-3B	Ministral-8B
Generated	3,400	3,443	3,400
QAs	1,042	685	861
Success	30.6	19.9	25.3

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

1648 G Detailed Dataset Characteristics

1649 Table 9 presents the distribution of reasoning types
1650 in our dataset. Following Tang and Yang (2024)’s
1651 taxonomy, we categorise questions based on their
1652 reasoning requirements into three fundamental
1653 types, yet given the complexity of multi-hop ques-
1654 tions, we consider their combinations as well. The
1655 majority of questions focus on inference reasoning
1656 in our dataset, requiring the derivation of logical
1657 conclusions by synthesising information across re-
1658 trieved chunks, C_q . The most common complex
1659 pattern combines inference and comparison rea-
1660 soning, while 9.8% of questions integrate all three
1661 reasoning types.

1662 Beyond reasoning types, our benchmark also
1663 varies in information synthesis requirements.
1664 While 47.6% of questions require synthesising
1665 information from two chunks, a substantial por-
1666 tion demands more complex integration: 19.4%
1667 require three chunks, 15.8% require four chunks,
1668 and 17.1% require five chunks. This distribution
1669 ensures our benchmark evaluates different complex
1670 information synthesis capabilities.

1671 H RAG Implementation Details for 1672 Reproducibility

1673 H.1 Model Sources

1674 We list all the model versions we used in our exper-
1675 iments below in the table 10.

1676 H.2 Model Parameters

1677 For reproducibility purpose, we list language model
1678 parameters in the table 11.

1679 H.3 Computational Resources

1680 All the VERGE experiments were conducted on
1681 a MacBook Pro equipped with an Apple M3 Max
1682 processor and 48 GB of unified memory. GPU-
1683 accelerated computations were performed using

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

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)
Minstral-8B	2024-10	bartowski/Minstral-8B-Instruct-2410-GGUF (Minstral-8B-Instruct-2410-Q4_K_M.gguf)
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

Table 10: List of LLMs used in our experiments

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.

I RAG results details

Table 12 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 13 demonstrates the RAG results only with CoT prompting strategy with its number of questions in the dataset.

J Ablation Study on Complete Multi-hop

Table 14 presents the complete ablation study results on multi-hop questions.

K Error Category Details

K.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 15 demonstrates the definition of each subcategory.

K.2 Error Category Examples

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

L Open-ended Question Answer Generation

L.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:

- Questions must necessitate synthesis across all provided documentary chunks

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

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) 1.28×10^7
- B) 2.27×10^8
- C) 7.70×10^{10}
- D) 3.37×10^7

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: The 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.

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
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

Table 11: Hyperparameters of the models.

Retriever	Chunk	Gemma2-9B		Llama3.1-8B		Ministral-8B	
		Standard	CoT	Standard	CoT	Standard	CoT
Closed	-	52.2	55.3	52.4	54.6	49.3	53.6
		(2,258)	(2,238)	(2,258)	(1,845)	(2,258)	(2,196)
Oracle	4000	69.7	73.9	64.8	69.9	66.0	68.7
		(2,258)	(2,200)	(2,258)	(1,726)	(2,258)	(2,130)
Dense	512	63.0	64.6	60.7	59.0	58.0	57.9
		(2,258)	(2,071)	(2,258)	(1,252)	(2,258)	(2,127)
	2048	64.7	66.8	60.4	64.7	57.1	60.8
	(2,258)	(2,124)	(2,258)	(1,746)	(2,258)	(2,116)	
	4000	63.6	68.4	58.8	68.2	57.9	62.6
	(2,258)	(2,145)	(2,258)	(1,704)	(2,258)	(2,111)	
Rerank	512	63.0	64.0	60.9	58.2	58.3	57.9
		(2,258)	(2,061)	(2,258)	(1,245)	(2,258)	(2,117)
	2048	64.0	66.9	60.2	64.0	57.2	60.6
	(2,258)	(2,243)	(2,258)	(1,738)	(2,258)	(2,106)	
4000	63.1	68.6	59.2	67.2	58.4	63.8	
(2,258)	(2,147)	(2,258)	(1,746)	(2,258)	(2,097)		
Sparse	512	61.1	63.5	60.4	58.5	58.6	55.6
		(2,258)	(1,921)	(2,258)	(1,131)	(2,258)	(2,086)
	2048	63.2	67.7	59.9	63.5	57.5	61.2
	(2,258)	(2,056)	(2,258)	(1,697)	(2,258)	(2,095)	
4000	63.1	69.1	59.2	66.4	59.0	63.7	
(2,258)	(2,096)	(2,258)	(1,682)	(2,258)	(2,099)		

Table 12: Accuracy in % with sample sizes in parentheses. Results across three LLMs, three chunk sizes (512, 2048, 4000), three retrievers, and two prompting strategies (Standard, CoT: Chain-of-Thought), plus closed-book and Oracle settings. The best configurations per model and prompt are indicated by bold.

1722 1723	2. Answer derivation must be impossible using only subset of chunks	<ul style="list-style-type: none"> • Synthesis: “Construct an argument for/against based on all provided information” 	1736 1737
1724 1725	3. Reference answers must demonstrate explicit multi-hop reasoning chains	<ul style="list-style-type: none"> • Evaluative: “Evaluate effectiveness of X considering mentioned constraints” 	1738 1739
1726 1727	4. Questions should elicit analytical, comparative, or causal reasoning	Multi-hop Reasoning Patterns:	1740
1728	Targeted Question Categories:	1. Chain Reasoning: Information(Chunk A) → Understanding(B) → Explanation(C)	1741 1742
1729 1730 1731	<ul style="list-style-type: none"> • Analytical: “Analyze the relationship between X and Y based on provided information” 	2. Convergent Reasoning: Multiple chunks provide complementary aspects converging to unified conclusion	1743 1744 1745
1732 1733	<ul style="list-style-type: none"> • Comparative: “Compare and contrast approaches described in documentation” 	3. Comparative Integration: Contrasting information across chunks to derive insights	1746 1747
1734 1735	<ul style="list-style-type: none"> • Causal: “Explain how X leads to Y, considering all mentioned factors” 	4. Conditional Dependencies: If-then relationships spanning multiple documentary sources	1748 1749

Retriever	Chunk	Gemma2-9B	Llama3.1-8B	Ministral-8B
Closed	-	54.8	44.6	52.1
Oracle	4000	72.0	53.4	64.8
Dense	512	59.3	32.7	54.6
	2048	62.8	50.0	57.0
	4000	65.0	51.5	58.6
Rerank	512	58.5	32.1	32.7
	2048	59.1	49.3	56.5
	4000	65.2	52.0	59.3
Sparse	512	54.5	29.3	54.6
	2048	61.6	47.7	56.8
	4000	64.1	49.4	59.2

Table 13: Accuracy in % with sample sizes in parentheses. Results across three LLMs, three chunk sizes (512, 2048, 4000), three retrievers, and two prompting strategies (Standard, CoT: Chain-of-Thought), plus closed-book and Oracle settings. The sample size is 2,258 for all the experiments.

	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 14: 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.

1750	Input Parameters:	L.2 Open-ended Question Answer	1771
1751	Domain: {task_domain}	Verification Prompt	1772
1752	Documentation (all chunks mandatory for utilization):	<hr/>	
1753	{documentation}	Multi-hop Reasoning Verification Framework	1773
1754		Objective:	1774
1755	Output Format Specification:	Perform comprehensive verification of multi-hop reasoning requirements for open-ended questions through systematic analysis.	1775
1756	Question: [Complex open-ended question requiring multi-hop reasoning]	Verification Dimensions:	1777
1757	Reference Answer: [Comprehensive answer demonstrating information synthesis from all chunks, minimum 3-5 sentences]	1. Multi-hop Reasoning Verification	1778
1758	Reasoning Steps: [Key reasoning process steps]	<ul style="list-style-type: none"> Determine necessity of cross-chunk information synthesis 	1779
1759	Quality Criteria:	<ul style="list-style-type: none"> Identify all chunks required for complete answer formulation 	1780
1760	<ul style="list-style-type: none"> Requires deep understanding and synthesis of all chunks 	<ul style="list-style-type: none"> Detect potential reasoning shortcuts bypassing multi-hop requirements 	1781
1761	<ul style="list-style-type: none"> Contains clear, comprehensive reference answer 	2. Answer Completeness Assessment	1782
1762	<ul style="list-style-type: none"> Prevents shortcut solutions or partial information usage 	<ul style="list-style-type: none"> Verify comprehensive synthesis of required information 	1783
1763	<ul style="list-style-type: none"> Encourages sophisticated reasoning processes 	<ul style="list-style-type: none"> Validate explicit multi-hop reasoning demonstration 	1784
1764			1785
1765			1786
1766			1787
1767			1788
1768			1789
1769			1790
1770	Generate only specified output format		

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 15: Definitions of each error category

1791	• Ensure answer non-derivability from chunk subsets	"confidence": integer (1-5 scale)	1821
1792		}	1822
1793	3. Chunk Utilization Analysis	Critical Constraints:	1823
1794	• Confirm essentiality of all provided chunks	Ensure question non-answerability through:	1824
1795	• Identify insufficiently integrated chunks	• Partial chunk utilization	1825
1796	• Assess cross-chunk integration depth	• General knowledge without documentary evidence	1826
1797	4. Improvement Recommendations	• Surface-level reading without synthesis	1827
1798	• Multi-hop failure: Suggest comprehensive chunk incorporation strategies		1828
1799		Output JSON response exclusively	1829
1800	• Answer incompleteness: Recommend completeness-enhancing details		
1801		L.3 Open-ended Question Answer Regeneration Prompt	1830
1802	• Chunk underutilization: Propose integration enhancement methods		1831
1803			
1804	Input Variables:	Multi-hop Question Regeneration Protocol	1832
1805	Question to Verify: {question}	Objective:	1833
1806	Reference Answer: {reference_answer}	Regenerate open-ended questions to satisfy comprehensive multi-hop reasoning requirements based on verification feedback.	1834
1807	Source Documents: {chunk_text}		1835
1808	Output JSON Schema:		1836
1809	{	Regeneration Context:	1837
1810	"required_chunks": [list of essential chunk indices],	Original Question: {original_question}	1838
1811	"reasoning": {	Original Answer: {original_answer}	1839
1812	"multi_hop_required": boolean,	Verification Feedback:	1840
1813	"answer_completeness": boolean,	• Unused Chunks Requiring Incorporation:	1841
1814	"shortcut_possible": boolean,	{unused_chunks}	1842
1815	"unused_chunks": [list of	• Specific Improvement Requirements: {suggestions}	1843
1816	underutilized chunk indices],		1844
1817	"improvement_suggestions": string		
1818	(mandatory)	Regeneration Requirements:	1845
1819	},		
1820	}		

1846	1. Mandate information utilization from all provided chunks	legal challenges related to due process violations.	1886
1847		The documentation highlights several key issues,	1887
1848	2. Prohibit answerability using chunk subsets	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.	1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900
1849	3. Ensure clear, comprehensive reference answer formulation		
1850			
1851	4. Demonstrate genuine multi-hop reasoning pathways		
1852			
1853	5. Eliminate shortcuts and surface-level reasoning possibilities		
1854			
1855	Input Parameters:		
1856	Domain: {task_domain}		
1857	Documentation:		
1858	{chunk_text}		
1859	Output Format Specification:		
1860	Question: [Improved question requiring all chunks]		
1861	Reference Answer: [Comprehensive answer synthesizing all chunks, 3-5 sentences]		
1862	Improvements Made: [Brief description of feedback incorporation]		
1863			
1864			
1865			
1866	Quality Assurance Criteria:		
1867	• Complete feedback incorporation		
1868	• All chunk utilization verification		
1869	• Multi-hop reasoning pathway clarity		
1870	• Answer comprehensiveness validation		
1871	Generate improved question addressing all feedback		
1872			
1873	<hr/>		
1874	L.4 Open-ended Question Answer Example		
1875	L.4.1 Example GovReport Question		
1876	<hr/>		
1877	Question:		
1878	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?		
1879			
1880			
1881			
1882			
1883			
1884	Reference Answer: Students accused of sexual misconduct in public universities face significant		
1885			
		L.4.2 Example HotpotQA Question	1901
		<hr/>	1902
		Question:	1903
		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?	1904 1905 1906 1907 1908 1909 1910
		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.	1911 1912 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927
		<hr/>	
		L.4.3 Example MultiFieldQA Question	1928
		<hr/>	1929
		Question:	1930
		Analyze the authors' approach to dialect identi-	1931

1932 fication in the context of the challenges and ad-
1933 vancements discussed in the provided documenta-
1934 tion. How does their two-stage pipeline address
1935 the complexities of distinguishing between similar
1936 languages and dialects, and what evidence do they
1937 provide to support its effectiveness?

1938 **Reference Answer:** The authors propose a two-
1939 stage pipeline for dialect identification, consisting
1940 of a language identification (LID) model followed
1941 by language-specific dialect classification models.
1942 This approach addresses the challenge of distin-
1943 guishing between similar languages and dialects
1944 by first accurately identifying the language of the
1945 input text, then leveraging language-specific mod-
1946 els for more precise dialect classification. They
1947 demonstrate the effectiveness of their pipeline by
1948 achieving top performance in both tracks of the
1949 VarDial 2023 DSL-TL shared task, outperforming
1950 other participants and previous works. Their re-
1951 sults, including a 58.54% accuracy in Track-1 and
1952 85.61% in Track-2, provide strong evidence for the
1953 superiority of their two-stage approach.

1954 L.4.4 Example SecFilings Question

1956 **Question:**
1957 How does PetroCorp’s approach to asset mainte-
1958 nance and drilling strategies, as described in their
1959 forward-looking statements, aim to mitigate the
1960 inherent volatility of oil and gas markets, particu-
1961 larly the fluctuations in prices, as outlined in their
1962 risk factors? **Reference Answer:** PetroCorp ac-
1963 knowledges the volatile nature of the oil and gas
1964 market, recognizing that prices are susceptible to
1965 fluctuations due to supply and demand, market un-
1966 certainty, and external factors beyond their control.
1967 To mitigate these risks, the company emphasizes
1968 a two-pronged approach. Firstly, PetroCorp pri-
1969 oritizes maintaining its property, plant, and equip-
1970 ment in good operating condition, as stated in their
1971 forward-looking statements. This proactive asset
1972 management strategy aims to minimize operational
1973 disruptions and ensure efficient production, thereby
1974 reducing the impact of price volatility on their op-
1975 erations. Secondly, PetroCorp emphasizes careful
1976 planning and execution of drilling activities, aim-
1977 ing to optimize production while acknowledging
1978 the uncertainties inherent in estimating reserves
1979 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 max-
imizing its operational efficiency and profitability.

L.4.5 Example 2WikiMultiHopQA Question

Question:
Analyze the evolution of Wheatstone’s contribu-
tions to communication technology, considering
both his successes and the controversies surround-
ing 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 mea-
surement 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 com-
plexities of invention and recognition in a rapidly
evolving technological landscape.