NOVELHOPQA: Diagnosing Multi-Hop Reasoning Failures in Long Narrative Contexts

Anonymous ACL submission

Abstract

Current large language models (LLMs) struggle to answer questions that span tens of thousands of tokens, especially when multi-hop reasoning is involved. While prior benchmarks explore long-context comprehension or multi-hop reasoning in isolation, none jointly vary context length and reasoning depth in natural narrative settings. We introduce **NOVELHOPQA**, the first benchmark to evaluate 1–4 hop QA over 64k–128k-token excerpts from 83 full-length public-domain novels. A keyword-guided pipeline builds hop-separated chains grounded in coherent storylines. We evaluate six state-of-the-art (SOTA) models and apply golden context filtering to ensure all questions are genuinely answerable. Human annotators validate both alignment and hop depth. We noticed **consistent accuracy drops with increased hops and context length**, even in frontier models—revealing that sheer scale does not guarantee robust reasoning. Our failure mode analysis highlights common breakdowns, such as missed final-hop integration and long-range drift. **NOVELHOPQA** offers a controlled diagnostic setting to stress-test multi-hop reasoning at scale.

1 Introduction

Understanding a question whose answer is scattered across tens of thousands of tokens is still beyond today's language models. Readers, lawyers, and historians trace clues across entire corpora, yet current NLP systems remain tuned to snippets only a few paragraphs long. When crucial evidence is buried in the middle of a long context, accuracy can plunge by more than 20 points (Liu et al., 2023a). Even frontier models score below 50% exact match on multi-document suites such as FanOutQA—where each query spans several Wikipedia pages—showing that larger context windows alone cannot solve cross-document reasoning (Zhu et al., 2024).

In principle, multi-hop benchmarks should reveal this weakness, but existing resources split into two groups. WikiHop and HotpotQA probe two-hop reasoning yet restrict inputs to a page or less (Welbl et al., 2018; Yang et al., 2018). At the other extreme, long-form sets such as NarrativeQA, QuALITY, NovelQA, and NoCha embrace book-scale inputs but ask mostly single-hop or summary questions (Kočiský et al., 2017; Pang et al., 2022; Wang et al., 2024a; Karpinska et al., 2024). Stress tests like MuSiQue's compositional traps and BABILong's million-token haystacks further highlight positional brittleness but rely on stitched or synthetic text (Trivedi et al., 2022; Kuratov et al., 2024).

Standardized long-context suites—LongBench, LEval, RULER, Marathon—show that models use a fraction of their window sizes while keeping hop depth fixed (Bai et al., 2024; An et al., 2023; Hsieh et al., 2024; Zhang et al., 2024). They do not reveal how context length interacts with reasoning depth.

Architectural advances offer partial relief. Sparse-attention models such as Longformer and BigBird reach 16–32 k tokens (Beltagy et al., 2020; Zaheer et al., 2021); recurrence and compression extend reach still further (Wu et al., 2022); and rotary extensions break the 100 k-token barrier (Ding et al., 2024). Yet retrieval-augmented or attribution-guided pipelines continue to outperform context-only baselines even at 32 k+ tokens (Xu et al., 2024; Li et al., 2024c). No public dataset simultaneously varies *(i) hop depth* and *(ii) authentic narrative context* $\geq$ *64k tokens*, preventing a principled diagnosis of long-context failures.

Existing benchmarks rarely test multi-hop reasoning over long, natural context. So we ask: **can models perform multi-step reasoning across 64k–128k tokens?** We introduce **NOVELHOPQA**, the first benchmark to jointly vary hop count (1–4) and narrative length, built from 83 novels with four balanced 1,000-example splits.

Contributions

(1) Public benchmark: 4,000 multi-hop QA examples spanning 64k–128k-token contexts.

(2) Reproducible pipeline: open-sourced extraction and paragraph-chaining code.

(3) Human validation: ten annotators confirm high alignment ($> 6.5/7$) and hop-match accuracy ($> 94\%$), ensuring dataset quality.

(4) Empirical hop-depth study: evaluations on six SOTA models trace accuracy decay along both axes.

Simply enlarging windows is *necessary but not sufficient*; true progress on long-context multi-hop reasoning demands benchmarks like **NOVELHOPQA** that stress both length and depth.

2 Related Work

Architectural, retrieval, and memory methods for long contexts. To process longer inputs, sparse-attention and recurrence-based architectures—Longformer, BigBird, Transformer-XL, and LongRoPE—scale attention and positional encodings to tens or hundreds of thousands of tokens (Beltagy et al., 2020; Zaheer et al., 2021; Dai et al., 2019; Ding et al., 2024). Retrieval-augmented generation and external-memory approaches boost performance when evidence is scattered (Lewis et al., 2021; Wu et al., 2022). Stress-test challenges like "Lost in the Middle" and NeedleBench highlight positional and retrieval brittleness in passages (Liu et al., 2023b; Li et al., 2024b), while BABILong probes reasoning limits with synthetic million-token haystacks (Kuratov et al., 2024). Although these advances surface key failure modes, they do not explore how reasoning depth interacts with very long contexts in natural prose.

Multi-hop QA benchmarks. WikiHop and HotpotQA pioneered cross-document and two-hop reasoning over short Wikipedia passages, with HotpotQA adding annotated supporting facts for explainability (Welbl et al., 2018; Yang et al., 2018). These datasets catalyzed advances in multi-hop inference but restrict inputs to at most a few thousand tokens—far from book-length scales. Subsequent compositional benchmarks such as MuSiQue introduce three-hop questions and trap-style tests (Trivedi et al., 2022), yet still operate on synthetic or stitched contexts rather than continuous narratives.

Long-context QA benchmarks. NarrativeQA and QuALITY probe book- or script-length inputs but mostly ask summary questions (Kočiský et al., 2017; Pang et al., 2022). NoCha and NovelQA raise the ceiling to 200k tokens, with NovelQA including both single- and multi-hop questions grounded in narrative detail (Wang et al., 2024a; Karpinska et al., 2024). More recent datasets expand the scope further: LooGLE controls for training-data leakage while comparing short- and long-dependency reasoning over 24k+ token documents (Li et al., 2024a); LV-Eval adds five length bands up to 256k tokens and misleading facts to test robustness (Yuan et al., 2024); and Loong focuses on multi-document QA with inputs drawn from domains like finance, law, and academia, frequently exceeding 100k tokens (Wang et al., 2024b). FanOutQA complements these length-centric benchmarks by evaluating reasoning breadth across multiple Wikipedia pages (Zhu et al., 2024). However, none of these benchmarks simultaneously test reasoning depth and long-context comprehension in coherent narratives—an issue that **NOVELHOPQA** addresses.

3 Dataset Construction

We build **NOVELHOPQA**—a benchmark that probes reasoning over book-length contexts (64k–128k tokens) with hop depths $H \in \{1, 2, 3, 4\}$. The pipeline comprises four stages: **(1)** novel selection, **(2)** anchor–keyword discovery, **(3)** paragraph chaining *with incremental QA generation*, and **(4)** final QA validation. *After each hop*, we regenerate the QA pair to integrate the newly appended paragraph, so the final 4-hop item reflects four rounds of question refinement rather than a single pass at the end.

3.1 Source Corpus

We selected 83 English novels from Project Gutenberg[1] (Gutenberg, 2025), a widely used repository of digitized books. We initially hand chose 100 diverse novels across genres and filtered this set down to 83 by removing books with fewer than 128k tokens after preprocessing. The final selection spans mystery, adventure, romance, and literary classics; includes both first- and third-person narration.

[1] https://www.gutenberg.org — All texts are in the U.S. public domain and legally permitted for research and redistribution. Our dataset annotations and processing code are released under the CC-BY-SA-4.0 license.

Hop	o1	4o	4o-mini	Gemini 2.5 P	Gemini 2.0 F	Gemini 2.0 FL	Avg.
1	95.90	95.60	92.30	**96.80**	93.10	90.90	94.10
2	95.50	95.40	91.80	**96.50**	92.80	90.30	93.72
3	95.20	95.10	91.30	**96.30**	92.40	90.00	93.38
4	94.80	94.90	90.90	**96.20**	92.10	89.60	93.08
Avg.	95.35	95.25	91.58	**96.45**	92.60	90.20	93.57

Table 1: Accuracy (%) of each model on **NOVELHOPQA** when evaluated using the original golden context.

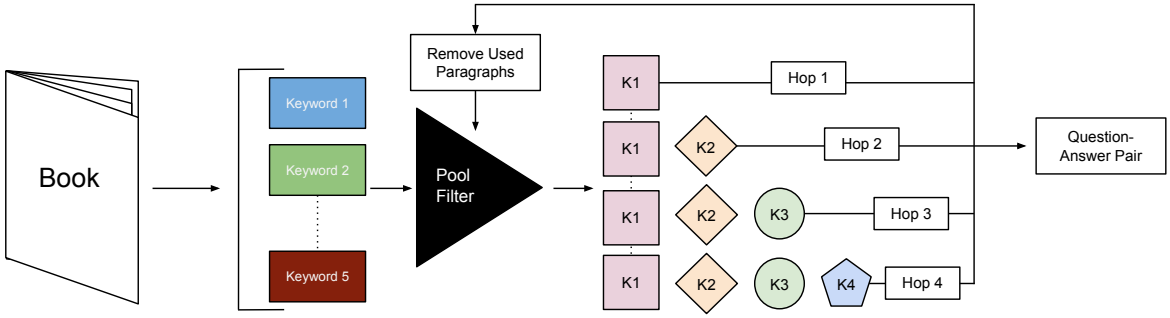

Figure 1: Keyword-guided paragraph-chaining pipeline used to build **NOVELHOPQA**. See Appendix F for a full example showing multi-hop evolution across four refinement stages.

3.2 Salient Keyword Filtering

For each of the 83 novels, we prompt GPT-4o-mini (OpenAI, 2024a) to suggest five "anchor" keywords—characters, locations, or objects central to the plot (see Appendix H for prompt). If any keyword appears fewer than 50 times in the text, we discard and re-sample that anchor, repeating up to seven times to ensure five high-frequency anchors.

3.3 Paragraph Pool Creation

We split each novel at blank lines and discard paragraphs under 30 words. The remaining paragraphs form a sampling pool for context construction.

3.4 Multi-Hop Context Chaining & Incremental QA Generation

For each book and hop depth $H \in \{1, 2, 3, 4\}$, we assemble contexts and QA pairs as follows (see Appendix H for all prompts):

1. **Hop 1:** Select a paragraph containing one of the book's anchor keywords k_1. Prompt GPT-4o (OpenAI, 2024a) to generate a single-hop QA pair (Q_1, A_1) from this paragraph.
2. **Hops** $h \in \{2 - H\}$**:**
   (a) Extract a new keyword k_h from the context C_{h-1} using our related-keyword prompt.
   (b) Sample a paragraph that contains both k_1 and k_h, and append it to the growing context $C_h = C_{h-1} \,\|$ new-paragraph.
   (c) Prompt GPT-4o to *re-generate* a single QA pair (Q_h, A_h) over the full context C_h, making sure the new QA integrates evidence from all h paragraphs.
3. **Paragraph exclusivity:** Remove each selected paragraph from the pool to prevent reuse. If no matching paragraph is found after seven attempts, abort the chain and restart with a fresh anchor.

This process "matures" each datapoint from (C_1, Q_1, A_1) through (C_H, Q_H, A_H), yielding naturally coherent multi-hop QA examples grounded in authentic narrative contexts.

3.5 Golden Context Filtering

To build a cleaner dataset focused on long-context reasoning, we evaluate all six models on the original golden context for each QA pair. As shown in Table 1, all models score above 90% on average, confirming that the questions are answerable. We remove any question missed by any model. Appendix 5 reports how many were removed per hop.

3.6 Fake and No Context Sanity Check

To confirm that questions require real context, we evaluated 800 QA pairs—100 per hop—under fake and no context settings. Accuracy stayed low, showing the questions aren't answerable without meaningful input. This helps ensure the dataset reflects reasoning, not recall. Full results are in Appendix E.

Metric	$H = 1$	$H = 2$	$H = 3$	$H = 4$
Alignment (1–7)	6.69	6.58	6.58	6.57
Hop Match (%)	95.9	94.9	94.9	95.2

Table 2: Average human validation scores across hop depths $H \in \{1, 2, 3, 4\}$. Alignment is the mean Likert score (1–7); Hop Match is the percentage judged to require exactly H steps. See Appendix C for full table.

Context	Hop	o1	4o	4o-mini	Gemini 2.5 P	Gemini 2.0 F	Gemini 2.0 FL	Avg.
64k	1	**92.51**	90.12	75.49	92.34	87.37	82.53	86.73
	2	87.66↓4.85	84.25↓5.87	74.77↓0.72	**87.84**↓4.50	77.02↓10.35	71.39↓11.14	80.48↓6.25
	3	84.99↓2.67	81.34↓2.91	73.14↓1.63	**85.12**↓2.72	74.25↓2.77	70.05↓1.34	78.13↓2.35
	4	82.15↓2.84	78.47↓2.87	68.04↓5.10	**82.45**↓2.67	71.76↓2.49	65.33↓4.72	74.69↓3.44
96k	1	**90.35**	88.83	72.25	90.12	82.26	78.44	83.71
	2	85.88↓4.47	82.67↓6.16	67.44↓4.81	**86.03**↓4.09	74.02↓8.24	67.04↓11.40	77.18↓6.53
	3	83.41↓2.47	80.41↓2.42	66.97↓0.47	**83.71**↓2.32	73.38↓0.64	66.05↓0.99	75.66↓1.52
	4	80.68↓2.73	76.92↓3.91	65.59↓1.38	**80.98**↓2.73	70.26↓3.12	62.81↓3.24	72.87↓2.79
128k	1	88.76	86.95	70.03	**89.10**	81.77	75.31	81.99
	2	84.33↓4.43	80.52↓6.43	63.95↓6.08	**84.70**↓4.40	69.13↓12.64	62.21↓13.10	74.14↓7.85
	3	81.92↓2.41	78.03↓2.92	62.95↓1.00	**82.20**↓2.50	68.78↓1.35	62.07↓0.14	72.66↓1.48
	4	**78.80**↓3.12	74.64↓3.31	61.18↓1.77	78.55↓3.65	67.32↓1.46	57.39↓4.68	69.65↓3.01

Table 3: Accuracy (%) on **NovelHopQA** across context lengths and hop depths, with mean performance in the last column. Red ↓ indicates drop from the previous hop; bold indicates the row-wise maximum. All cells with accuracy drops are highlighted in red. More graphs are included in Appendix A to further visualize these trends.

4 Human Evaluation

Ten undergraduate validators each annotated 260 examples—40 from the 1- and 2-hop sets, and 90 from the 3- and 4-hop sets. They rated **Alignment**, measuring how well each QA pair matched its source context, and judged **Hop Match**, assessing whether the answer required exactly H reasoning steps. See Appendix C for detailed results and Appendix G for the evaluation form.

5 Results and Discussion

We evaluate six models on **NovelHopQA** using chain-of-thought prompts: **o1** (OpenAI, 2024c), **Gemini 2.5 Pro** (DeepMind, 2025b), **GPT-4o** (OpenAI, 2024a), **GPT-4o-mini** (OpenAI, 2024a), **Gemini 2.0 Flash**, and **Gemini 2.0 Flash Lite** (DeepMind, 2025a). Table 3 summarizes model accuracy across three context lengths (64k, 96k, 128k) and four hop depths (1–4).

Impact of hop depth. All models exhibit consistent performance degradation as hop depth increases. On average, accuracy drops roughly 12 points from 1-hop to 4-hop at 64k context length. Even reasoning models like Gemini 2.5 Pro and o1 see steady declines with more complex multi-step questions, highlighting the challenge of compositional reasoning at scale.

Impact of context length. Longer context lengths also lead to reduced accuracy, though the effect is milder than that of hop count. Across models, 1-hop performance drops about 5 points when moving from 64k to 128k contexts. This suggests that deeper reasoning contributes more to failure than sheer length.

Model comparisons. Reasoning models—**Gemini 2.5 Pro** and **o1**—consistently outperform others, often topping each row in Table 3. Gemini 2.5 Pro achieves the highest average accuracy overall, followed closely by o1 and GPT-4o. Mid-sized models like GPT-4o-mini and Gemini Flash Lite perform noticeably worse, especially under 4-hop and 128k settings, where their scores fall into the 60s.

Robustness at scale. Despite large context windows, no model maintains strong performance on the hardest tasks (4-hop at 128k), where even top models dip below 80%. These results affirm that long-context capacity alone is not enough—robust multi-hop reasoning remains an open challenge. A deeper analysis as to why models fail is provided in Appendix D.

6 Conclusion

NovelHopQA is the first benchmark to vary both context length (64k–128k) and hop depth $H \in \{1, 2, 3, 4\}$ in long-context QA. Human validation confirms quality, and models show accuracy drops along both axes. These results highlight that **larger context windows aren't enough**—multi-hop reasoning remains a core challenge. Code and data will be released upon publication.

7 Limitations

NovelHopQA fills a key gap in long-context, multi-hop QA, but several limitations remain:

Genre and temporal bias. All contexts come from public-domain novels published before 1927, Project Gutenberg (Gutenberg, 2025). Their style, vocabulary, and themes reflect older literary English and omit modern genres (e.g., journalism, technical manuals) as well as non-literary domains. Including contemporary texts and non-English sources would improve representativeness.

Dialectal and domain diversity. Our data largely comprises standard literary English, with few regional or archaic dialects; LLM performance on non-standard varieties may differ substantially (Gupta et al., 2024, 2025).

Generation and grading bias. All QA pairs are generated by GPT-4o (OpenAI, 2024a), and correctness is automatically graded by GPT-4.1 (OpenAI, 2024b) with CoT prompts. Both steps risk inheriting model-specific patterns or blind spots. Human-authored questions and manual grading (or mixed human–machine adjudication) could reveal edge cases and reduce generator/grader artifacts.

Evaluation metric. We report accuracy as judged by GPT-4.1 (OpenAI, 2024b) using CoT evaluation prompts. This approach allows for some flexibility in phrasing and considers reasoning consistency. Future evaluations could incorporate human review or rationale-based scoring for more robust assessment.

8 Ethics Statement

Data provenance. All passages are sourced from public-domain novels on Project Gutenberg (Gutenberg, 2025). No private or sensitive data is included.

Annotator protocol. Ten undergraduate validators majoring in computer science, data science, or cognitive science (aged 18+) provided informed consent and were compensated for their time. They evaluated whether each question was answerable from its context, rated alignment, and verified that the reasoning depth matched the intended hop count (Table 6). No additional personal data were collected.

QA generation and grading. QA pairs were generated by GPT-4o (OpenAI, 2024a) and graded by GPT-4.1 (OpenAI, 2024b) using CoT prompting. To validate quality, human annotators assessed whether each question aligned with its context, whether it could be answered from the provided text, and whether the reasoning depth matched the intended hop count.

Intended use. **NovelHopQA** is provided for academic research on long-context, multi-hop reasoning. It is not intended for deployment in safety-critical or high-stakes applications without further validation.

Reproducibility Statement

We describe our dataset construction process in Section 3, and include all prompt templates in Appendix H. All model generations were obtained using publicly available APIs. Specifically, we used the Azure OpenAI API for GPT-4o, GPT-4o-mini (OpenAI, 2024a), and o1 (OpenAI, 2024c), and the Google Vertex API for Gemini 2.0 Flash, Flash Lite (DeepMind, 2025a), and Gemini 2.5 Pro (DeepMind, 2025b). All models were queried using CoT prompts, and their outputs were graded using GPT-4.1 (OpenAI, 2024b) with CoT-based evaluation prompts. We plan to release the dataset, prompts, and model outputs upon publication to support replication and further research.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *Preprint*, arXiv:2307.11088.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *Preprint*, arXiv:1901.02860.

Google DeepMind. 2025a. Gemini 2.0 flash and flash lite. Online documentation. Google Cloud, Vertex AI, and Google AI Studio documentation.

Google DeepMind. 2025b. Gemini model and thinking updates: March 2025. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-05-16.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *Preprint*, arXiv:2402.13753.

Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Austen Liao, Kevin Zhu, and Sean O'Brien. 2025. Endive: A cross-dialect benchmark for fairness and performance in large language models. *Preprint*, arXiv:2504.07100.

Abhay Gupta, Philip Meng, Ece Yurtseven, Sean O'Brien, and Kevin Zhu. 2024. Aavenue: Detecting llm biases on nlu tasks in aave via a novel benchmark. *Preprint*, arXiv:2408.14845.

Project Gutenberg. 2025. Project gutenberg. Accessed: 2025-04-17.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. *Preprint*, arXiv:2406.16264.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Preprint*, arXiv:1712.07040.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Preprint*, arXiv:2406.10149.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. Loogle: Can long-context language models understand long contexts? *Preprint*, arXiv:2311.04939.

Mo Li, , Songyang Zhang, Yunxin Liu, and Kai Chen. 2024b. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *Preprint*, arXiv:2407.11963.

Yanyang Li, Shuo Liang, Michael R. Lyu, and Liwei Wang. 2024c. Making long-context language models better multi-hop reasoners. *Preprint*, arXiv:2408.03246.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Introducing gpt-4.1 in the api. Accessed: 2025-05-17.

OpenAI. 2024c. Introducing openai o1. `https://openai.com/o1/`. Accessed: 2025-05-16.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. 2022. Quality: Question answering with long input texts, yes! *Preprint*, arXiv:2112.08608.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. 2024a. Novelqa: Benchmarking question answering on documents exceeding 200k tokens. *Preprint*, arXiv:2403.12766.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *Preprint*, arXiv:2406.17419.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Preprint*, arXiv:1710.06481.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. *Preprint*, arXiv:2203.08913.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. *Preprint*, arXiv:2310.03025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, Guohao Dai, Shengen Yan, and Yu Wang. 2024. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. *Preprint*, arXiv:2402.05136.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences. *Preprint*, arXiv:2007.14062.

Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. 2024. Marathon: A race through the realm of long context with large language models. *Preprint*, arXiv:2312.09542.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *Preprint*, arXiv:2402.14116.

A Breakdown Visualizations of Model Accuracy Trends

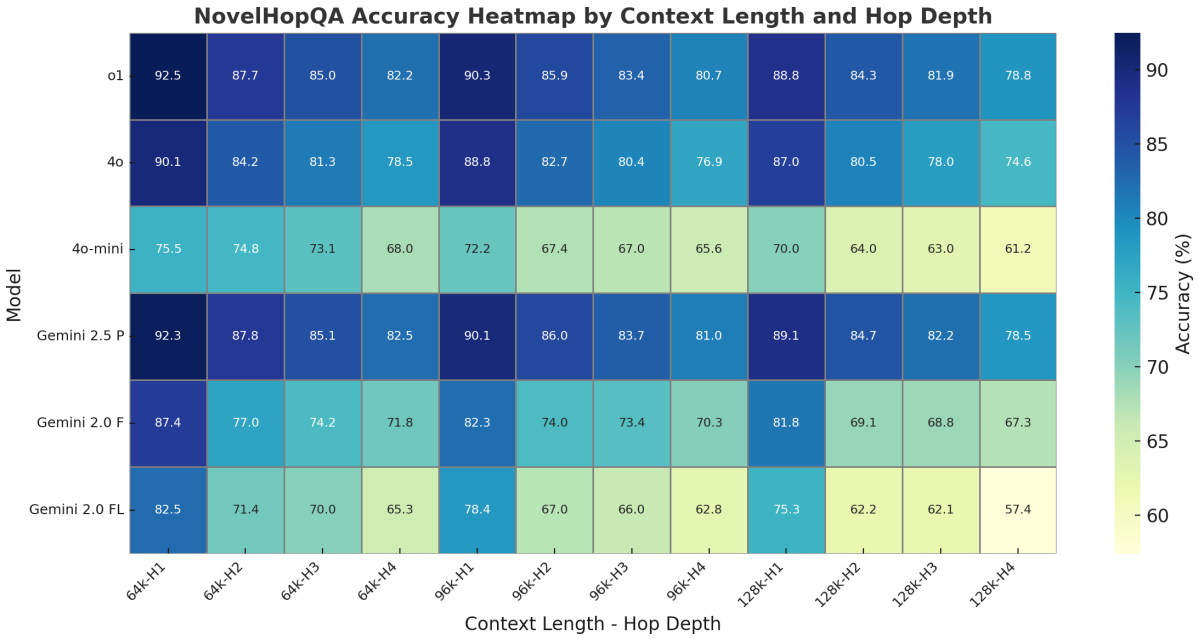

Figure 2: Accuracy (%) on **NovelHopQA** across context lengths and hop depths $H \in \{1, 2, 3, 4\}$. This heatmap shows how model accuracy declines as both narrative length and multi-hop reasoning depth increase.

To complement the heatmap, we include detailed line plots illustrating model-specific trends across each axis independently:

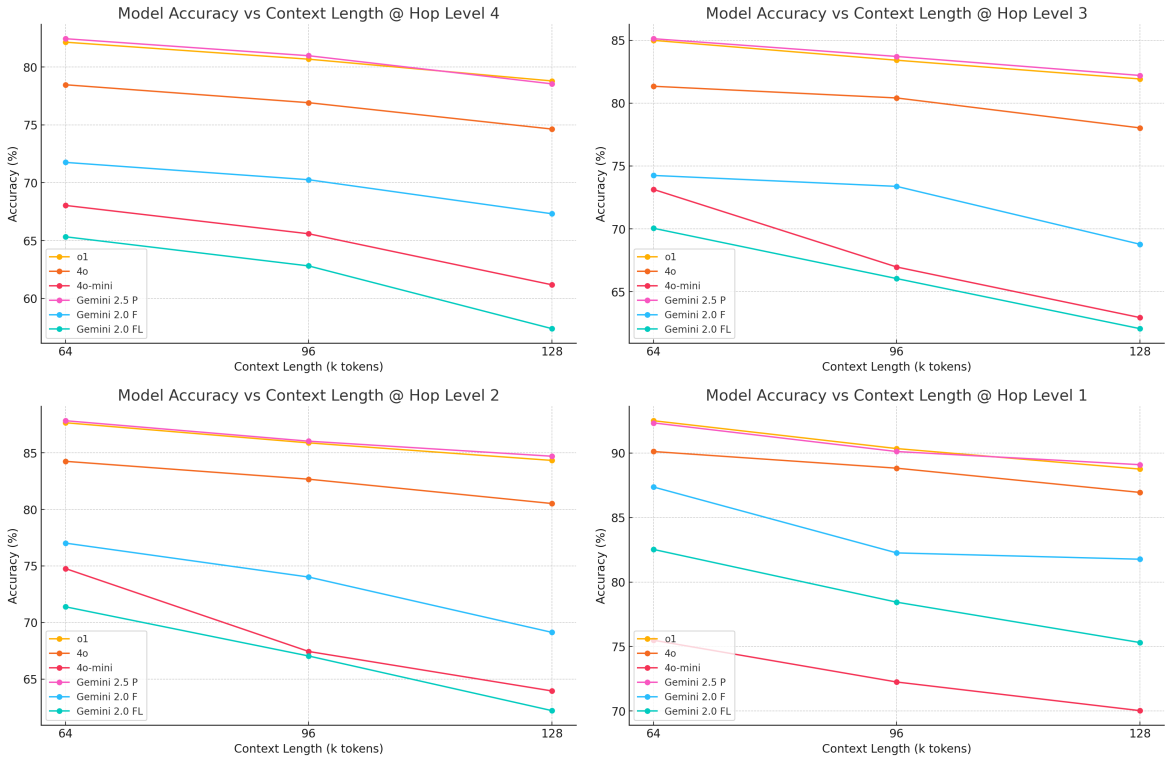

Figure 3: Model performance across context lengths for each hop level $H = 1, 2, 3, 4$. These plots isolate the effect of longer narratives on accuracy.

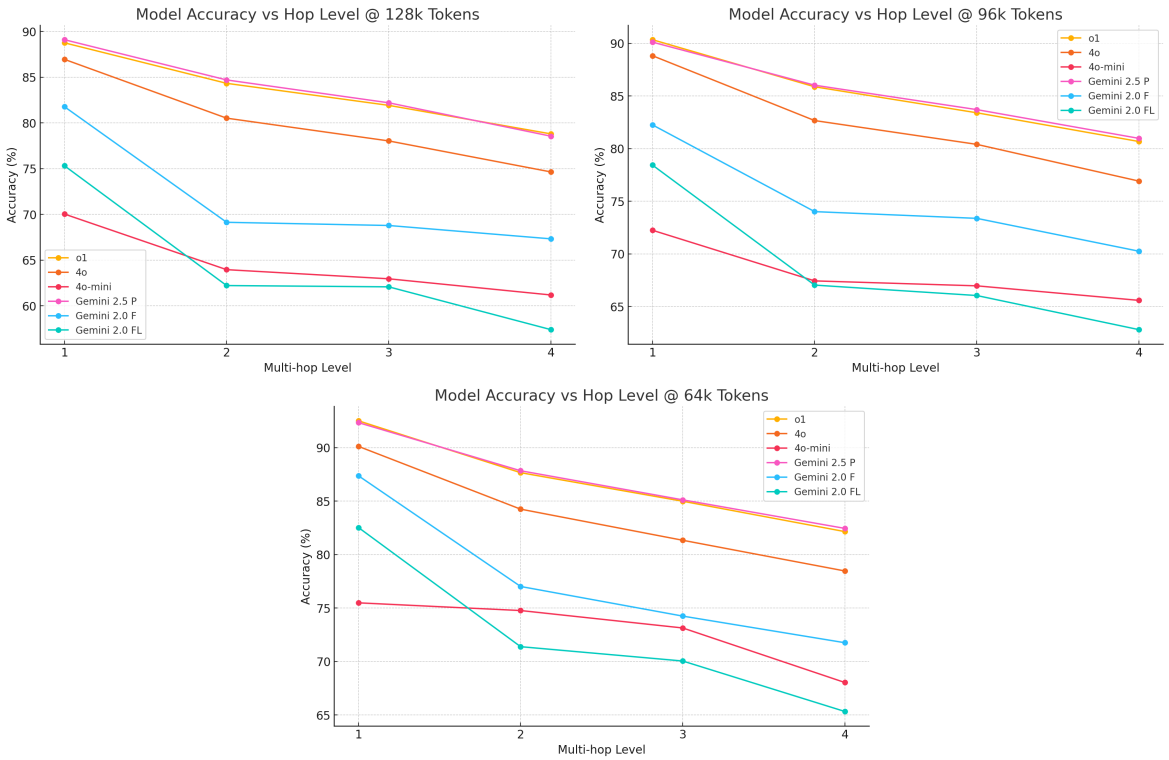

Figure 4: Model performance across hop levels for each context length (64k, 96k, 128k). These plots isolate the effect of deeper reasoning on accuracy.

B Dataset Statistics by Hop Level

Hop Level	Count	Avg. Context Tokens	Avg. Answer Length
1-Hop	1000	191.92	4.64
2-Hop	1000	451.46	6.99
3-Hop	1000	691.85	9.59
4-Hop	1000	916.82	10.79

Table 4: Dataset statistics across hop levels. Each row reports the number of QA pairs, the average context length in tokens, and the average answer length in words.

B.1 Filtered Dataset Size After Golden Context Evaluation

Hop Level	# Removed	New Total
1-Hop	37	963
2-Hop	39	961
3-Hop	40	960
4-Hop	42	958

Table 5: Number of questions removed per hop after golden context filtering.

C Full Human Evaluation Table

Validator	$H = 1$		$H = 2$		$H = 3$		$H = 4$	
	Align	**Hop Match**	**Align**	**Hop Match**	**Align**	**Hop Match**	**Align**	**Hop Match**
Validator 1	6.71	96.2	6.52	94.0	6.69	95.1	6.57	96.5
Validator 2	6.66	97.1	6.43	95.3	6.55	93.6	6.64	94.9
Validator 3	6.79	95.8	6.68	96.7	6.42	94.4	6.71	93.8
Validator 4	6.60	94.7	6.57	93.9	6.61	95.2	6.45	96.1
Validator 5	6.70	95.3	6.61	96.5	6.58	94.8	6.73	95.7
Validator 6	6.58	96.9	6.65	95.2	6.66	96.6	6.52	94.5
Validator 7	6.63	96.1	6.50	94.4	6.70	95.5	6.59	93.7
Validator 8	6.74	95.0	6.56	93.6	6.47	94.3	6.65	96.8
Validator 9	6.69	97.2	6.67	94.8	6.53	96.0	6.68	95.4
Validator 10	6.77	94.5	6.62	95.6	6.60	93.9	6.54	94.2
Average	**6.69**	**95.9**	**6.58**	**94.9**	**6.58**	**94.9**	**6.57**	**95.2**

Table 6: Full human validation scores across hop depths $H \in \{1, 2, 3, 4\}$. "Alignment" is the average Likert rating (1–7); "Hop Match" is the percentage of responses judged to require exactly H reasoning steps.

D Failure Mode Analysis

To analyze where models fail despite access to full 64k–128k token narratives in NOVELHOPQA, we identify four major reasoning failure types, each illustrated with concrete cases.

D.1 1. Missing Final-Hop Integration

Models retrieve evidence for early hops but omit the final inference step—often failing to incorporate a late-stage paragraph into their answer. This frequently occurs in 4-hop settings where the final clue appears deep in the context.

Hop	Question	Model Answer	Gold Answer
4	Which three combined revelations prompt Elizabeth to reassess Mr. Darcy?	Wickham's deceit and housekeeper's praise	Wickham's deceit, housekeeper's praise, and Georgiana's testimony

Table 7: Model retrieves early clues but omits Georgiana's late-paragraph testimony.

D.2 2. Entity Confusion / Coreference Errors

Ambiguous names or pronouns lead models to conflate characters or roles. This issue surfaces in dialogue-heavy novels or when entities share close proximity in the context.

Hop	Question	Model Answer	Gold Answer
3	Who secured Captain Ahab's rope at the rail before he was hoisted to his perch?	Queequeg	Starbuck

Table 8: Model confuses two nearby characters due to unclear attribution.

D.3 3. Incomplete Evidence Combination

When evidence is distributed across multiple hops, models sometimes answer using only part of the required chain—missing one or more critical supporting facts.

Hop	Question	Model Answer	Gold Answer
2	After reading Darcy's letter, which revelation begins to alter Elizabeth's opinion of Mr. Darcy?	Wickham squandered the inheritance	Wickham squandered the inheritance and attempted to elope with Georgiana

Table 9: Model recalls one clue but omits the additional elopement detail.

D.4 4. Contextual Drift

In very long contexts, models may focus on irrelevant segments while overlooking key evidence located far away. This issue arises most often in 128k-token contexts requiring long-range linkage between events.

Hop	Question	Model Answer	Gold Answer
4	What earlier interaction foreshadowed the villain's final betrayal in the court scene?	A vague warning by a servant	The servant's warning *and* the unsigned letter hidden in the drawer

Table 10: Model retrieves partial clue but misses distant supporting passage.

E Fake and No Context Evaluation

To evaluate whether models genuinely rely on the narrative context provided in **NOVELHOPQA**, we conduct an ablation study using two control conditions: **fake context** and **no context**. This analysis serves as a sanity check to verify that model accuracy is not attributable to memorization or dataset leakage.

Fake Context. For each question, we prompted the model with unrelated context. The paragraph has no semantic or lexical relationship to the QA pair. The fake context used is shown in Appendix Table 12.

No Context. The model is given only the question and no surrounding passage. This isolates performance that arises solely from model priors or memorized facts.

Experimental Setup. Each model was evaluated on 800 examples—100 random questions from each of four datasets, under both fake and no context conditions. All responses were graded by GPT-4.1 (OpenAI, 2024b) using CoT prompting for consistency.

Model	Condition	1-hop	2-hop	3-hop	4-hop
`Gemini 2.0 Flash Lite`	Fake context	4% (4/100)	3% (3/100)	1% (1/100)	1% (1/100)
	No context	4% (4/100)	3% (3/100)	1% (1/100)	1% (1/100)
`GPT-4o Mini`	Fake context	5% (5/100)	4% (4/100)	1% (1/100)	1% (1/100)
	No context	4% (4/100)	3% (3/100)	1% (1/100)	1% (1/100)
`Gemini 2.0 Flash`	Fake context	6% (6/100)	5% (5/100)	1% (1/100)	1% (1/100)
	No context	5% (5/100)	4% (4/100)	1% (1/100)	1% (1/100)
`GPT-4o`	Fake context	6% (6/100)	5% (5/100)	2% (2/100)	1% (1/100)
	No context	6% (6/100)	5% (5/100)	1% (1/100)	1% (1/100)
`o1`	Fake context	6% (6/100)	5% (5/100)	2% (2/100)	1% (1/100)
	No context	6% (6/100)	5% (5/100)	2% (2/100)	1% (1/100)
`Gemini 2.5 Pro`	Fake context	7% (7/100)	6% (6/100)	2% (2/100)	1% (1/100)
	No context	7% (7/100)	5% (5/100)	2% (2/100)	1% (1/100)

Table 11: Accuracy (%) on 100 randomly selected multi-hop questions under fake and no context settings. Models perform near chance across all hops, demonstrating that answers cannot be derived without relevant narrative input.

Fake Context Example (*The Secret Garden*)

Context Source:

1. **Paragraph 1.**
   It was the sweetest, most mysterious-looking place any one could imagine. The high walls which shut it in were covered with the leafless stems of climbing roses which were so thick that they were matted together. Mary Lennox knew they were roses because she had seen a great many roses in India. All the ground was covered with grass of a wintry brown, and out of it grew clumps of bushes which were surely rose-bushes if they were anything. There were numbers of standard roses which had so spread their branches that they were like little trees. There were other trees in the garden, and one of the things which made the place look strangest and loveliest was that climbing roses had run all over them and swung down long tendrils which made light swaying curtains.
2. **Paragraph 2.**
   And here and there among the grass were narcissus bulbs beginning to sprout and uncurl their narrow green leaves. She thought they seemed to be stretching out their arms to see how warm the sun was. She went from one part of the garden to another. She found many more of the sprouting pale green points and she found others which were white crocuses and snowdrops, because the green spikes had burst through their sheaths and showed white. She remembered what Ben Weatherstaff had said about the "snowdrops by the thousands," and about bulbs spreading and making new ones. "These had been left to themselves for ten years," perhaps, and they had spread like the snowdrops into thousands.

Table 12: The "fake context" passage used during ablation. This excerpt, unrelated to any QA pair, was paired with a question to test whether models output plausible answers.

Interpretation. This experiment validates the integrity of **NOVELHOPQA** by confirming that models are not simply memorizing QA pairs seen during pretraining. Accuracy remains near-zero when relevant context is removed, demonstrating that our questions are novel and context-dependent. These findings strengthen confidence that model performance on **NOVELHOPQA** reflects actual reading comprehension and not artifact exploitation or memorization.

F 4-Hop QA Evolution Example

4-Hop QA Evolution Example

Context Source (*Pride and Prejudice*):

1. **Paragraph 1 (Meryton Assembly).**
   "*She is tolerable, but not handsome enough to tempt me,*" Darcy said coldly, and Elizabeth—within earshot—coloured with mingled amusement and disdain.
2. **Paragraph 2 (Darcy's Letter).**
   In his long letter, Darcy disclosed that Wickham squandered the inheritance meant for him and then tried to elope with Georgiana for her fortune.
3. **Paragraph 3 (Pemberley Visit).**
   At Pemberley, the housekeeper praised Darcy's generosity, and Georgiana greeted Elizabeth with shy warmth, eager to confirm her brother's good opinion of her.
4. **Paragraph 4 (Rosings Confrontation).**
   Lady Catherine warned Elizabeth that Darcy was "destined for a connection of higher consequence," accidentally revealing his unwavering attachment.

Questions Across Hops:

- **Hop 1:** Why does Elizabeth form an unfavourable first impression of Mr. Darcy at the Meryton assembly?
- **Hop 2:** After reading Darcy's letter, which revelation begins to alter Elizabeth's opinion of him?
- **Hop 3:** During her visit to Pemberley, what fresh evidence further reinforces Elizabeth's changing view of Mr. Darcy's character?
- **Hop 4 (final):** Which combined revelations—including Georgiana's testimony—prompt Elizabeth to reassess her original judgment of Mr. Darcy?

Final Answer (Hop 4):
Wickham's deceit, exposed in Darcy's letter, the housekeeper's praise of Darcy's generosity, and Georgiana's sincere affection together reveal Darcy's true integrity, persuading Elizabeth to overturn her initial prejudice.

Hop Reasoning Breakdown:

- **Hop 1 — First Impression**: Interprets Darcy's slight at the assembly.
- **Hop 2 — Hidden Truth**: Integrates revelations from Darcy's letter, discrediting Wickham.
- **Hop 3 — Character Witness**: Adds Pemberley testimony and Georgiana's behaviour as proof of Darcy's virtue.
- **Hop 4 — Synthesis**: Combines all prior evidence to explain Elizabeth's reassessment.

Table 13: 4-hop QA example showing the step-wise evolution of context, question, and reasoning.

G Human Evaluation Form Example

Human Evaluation Form (3-Hop)

Paragraph 1:
Now, inclusive of the occasional wide intervals between the revolving outer circles, and inclusive of the spaces between the various pods in any one of those circles, the entire area at this juncture, embraced by the whole multitude, must have contained at least two or three square miles. [. . .] Queequeg patted their foreheads; Starbuck scratched their backs with his lance; but fearful of the consequences, for the time refrained from darting it.

Paragraph 2:
But not a bit daunted, Queequeg steered us manfully; now sheering off from this monster directly across our route in advance; now edging away from that, whose colossal flukes were suspended overhead, while all the time, Starbuck stood up in the bows, lance in hand, pricking out of our way whatever whales he could reach. [. . .]

Paragraph 3:
"I will have the first sight of the whale myself,"—he said. [. . .] Then arranging his person in the basket, he gave the word for them to hoist him to his perch, **Starbuck** being the one who secured the rope at last; and afterwards stood near it. [. . .]

Question: Who was tasked with securing Captain Ahab's rope at the rail before he was hoisted to his perch to get the first sight of the whale?

Does this question require single-hop reasoning? (circle one)
Yes No

Rate alignment on a 7-point Likert scale (circle one):
1 – Completely unrelated, 2 – Mostly unrelated, 3 – Somewhat related, 4 – Moderately related, 5 – Strongly related, 6 – Very closely related, 7 – Perfectly aligned

Table 14: Example form used by validators to assess hop depth and contextual alignment.

H Prompt Templates

Anchor Keyword Generation

```
You are a literary analysis expert. Based solely on the book title "{book_title}", list five main keywords central to its plot. Ensure each keyword is concise (one or two words) and appears at least 50 times.
Answer format:

<keyword_result>
keyword1;
keyword2;
keyword3;
keyword4;
keyword5
</keyword_result>
```

Figure 5: Prompt for extracting five high-frequency anchor keywords from a book title

Single Hop Generation

```
You are an expert question generator. Given the paragraph below, generate one challenging question that requires understanding of this paragraph. Provide a concise answer.
Output format:

<question>Your question here</question>
<answer>Your concise answer here</answer>

Paragraph: {paragraph}
```

Figure 6: Prompt for generating a single-hop question from one paragraph.

Extract Related Keyword

```
You are an expert at extracting related keywords. From the paragraph below, identify a keyword strongly related to its content but different from "{current_keyword}". Return only the new keyword.
Output format:

<keyword>NEW_KEYWORD</keyword>

Paragraph: {paragraph}
```

Figure 7: Prompt for extracting a related keyword at hop h.

Generate Final Multi-Hop Question

```
You are an expert multi-hop question generator. Generate one question requiring integration across all provided paragraphs, and provide a concise answer.
Output format:

<question>Your multi-hop question here</question>
<answer>Your concise answer here</answer>

Context: {paragraph1}\n\n {paragraph2}...\n\n {paragraphH}
```

Figure 8: Prompt for generating the final multi-hop question over H paragraphs.