Ko-LongRAG: A Korean Long-Context RAG Benchmark Built with a Retrieval-Free Approach

Yongil Kim **Heuiyeen Yeen** **Hyeongu Yun** **Jinsik Lee**
LG AI Research
{yong-il.kim, heuiyeen214, hyeongu.yun, jinsik.lee}@lgresearch.ai

Abstract

The rapid advancement of large language models (LLMs) significantly enhances long-context Retrieval-Augmented Generation (RAG), yet existing benchmarks focus primarily on English. This leaves low-resource languages without comprehensive evaluation frameworks, limiting their progress in retrieval-based tasks. To bridge this gap, we introduce Ko-LongRAG, the first Korean long-context RAG benchmark. Unlike conventional benchmarks that depend on external retrievers, Ko-LongRAG adopts a retrieval-free approach designed around Specialized Content Knowledge (SCK), enabling controlled and high-quality QA pair generation without the need for an extensive retrieval infrastructure. Our evaluation shows that o1 model achieves the highest performance among proprietary models, while EXAONE 3.5 leads among open-sourced models. Additionally, various findings confirm Ko-LongRAG as a reliable benchmark for assessing Korean long-context RAG capabilities and highlight its potential for advancing multilingual RAG research.[1]

1 Introduction

The rapid advancements in long-context large language models (LLMs) significantly enhance their ability to process and comprehend extended texts, benefiting diverse applications such as information retrieval, document summarization, and question answering (Naveed et al., 2023; Achiam et al., 2023). In response, numerous benchmarks are developed to evaluate the effectiveness of Retrieval-Augmented Generation (RAG) configurations (Chen et al., 2024). However, existing benchmarks (Chen et al., 2024; Friel et al., 2024) predominantly focus on English, leaving low-resource languages without comprehensive RAG evaluation frameworks (Chirkova et al., 2024). Moreover, the lack of extensive knowledge bases and the scarcity of research tasks in non-English languages further complicate benchmark construction.

To address this issue, we propose **Ko-LongRAG**, a high-quality **Ko**rean **Long**-context **RAG** benchmark, along with a novel approach for generating RAG datasets without reliance on explicit retrieval settings. Unlike conventional benchmarks that rely on existing retrievers, Ko-LongRAG leverages a retrieval-free paradigm designed around Specialized Content Knowledge (SCK). SCK refers to domain-specific knowledge that facilitates the generation of meaningful tasks without the need for an extensive retrieval infrastructure. By segmenting the corpus into domain-specific clusters and generating question-answer pairs within these clusters based on document similarity, Ko-LongRAG effectively simulates retrieval-based scenarios while maintaining high relevance and contextual fidelity.

This methodology offers several advantages: (1) it eliminates the dependency on external retrievers, ensuring applicability in low-resource settings; (2) it maintains the integrity of retrieval-like evaluation by clustering highly similar documents; and (3) it ensures the benchmark's scalability across diverse domains. Moreover, Ko-LongRAG evaluates models on answerability, which measures their ability to derive accurate answers based on the given context, reflecting the model's effectiveness in utilizing retrieved documents in a RAG setting.

Through our experiments, we employ Ko-LongRAG to assess the Korean long-context RAG performance of various LLMs and demonstrate the benchmark's robustness and utility. Among the proprietary models, o1 (OpenAI, 2024b) demonstrates superior performance, while EXAONE 3.5 (An et al., 2024) excels among open-sourced models. Additionally, we observe a strong correlation between model size and performance, a decline in accuracy when models support shorter context lengths—highlighting the necessity of long-context

[1] The dataset is available at https://huggingface.co/datasets/LGAI-EXAONE/Ko-LongRAG.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17317–17329
November 4-9, 2025 ©2025 Association for Computational Linguistics

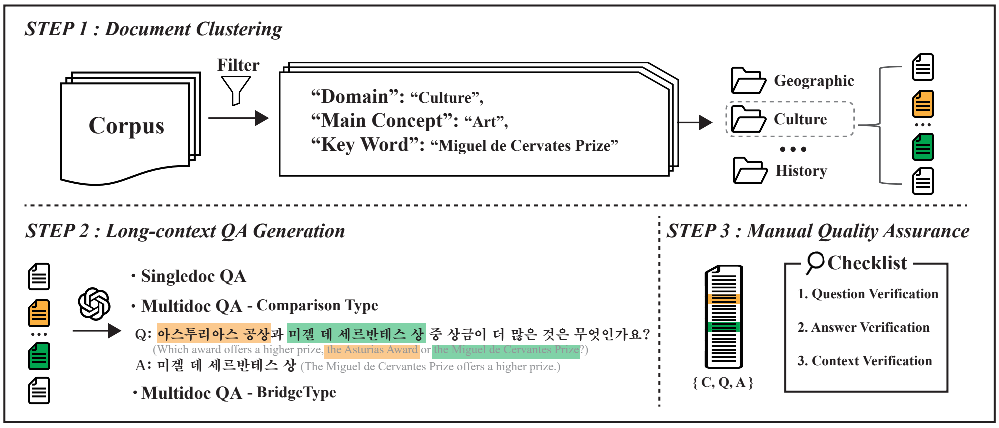

STEP 1 : Document Clustering
Corpus
Filter
"Domain": "Culture",
"Main Concept": "Art",
"Key Word": "Miguel de Cervates Prize"
Geographic
Culture
History
STEP 2 : Long-context QA Generation
· Singledoc QA
· Multidoc QA - Comparison Type
Q: 아스투리아스 공상과 미겔 데 세르반테스 상 중 상금이 더 많은 것은 무엇인가요?
(Which award offers a higher prize, the Asturias Award or the Miguel de Cervantes Prize?)
A: 미겔 데 세르반테스 상 (The Miguel de Cervantes Prize offers a higher prize.)
· Multidoc QA - BridgeType
STEP 3 : Manual Quality Assurance
{ C, Q, A }
Checklist
1. Question Verification
2. Answer Verification
3. Context Verification

Figure 1: Ko-LongRAG Construction Pipeline. The **C**, **Q**, **A** triplet (long-**C**ontext passage, **Q**uestion, **A**nswer) is created through SCK-based document clustering, LLM-based QA generation, and manual quality assurance.

processing for effective performance—and consistent robustness of results across various judge models. These findings validate Ko-LongRAG as an effective benchmark for Korean long-context RAG research and underscore its potential to drive innovation in multilingual RAG studies.

2 Related Works

To rigorously assess the evolving capabilities of long-context language models across dimensions such as comprehension, memory retention, and input scalability, a range of benchmarks has been proposed, including LongBench (Bai et al., 2023), Bamboo (Dong et al., 2023), Marathon (Zhang et al., 2023), and RULER (Hsieh et al., 2024). Similarly, in the context of retrieval-augmented generation (RAG), benchmarks such as RAGBench (Friel et al., 2024) and LongRAG (Jiang et al., 2024) have been introduced to evaluate multi-document reasoning under long-context settings. However, the majority of these benchmarks are developed for English (Chirkova et al., 2024), thereby limiting their applicability to low-resource languages.

In multilingual contexts, RAG evaluations typically rely on translated English datasets or multilingual LLMs, which inadequately capture language-specific retrieval challenges. In particular, Korean, in addition to lacking structured knowledge bases and retrieval infrastructures, also exhibits linguistic properties, such as morphological richness, agglutinative syntax, and flexible word order, that complicate token-level retrieval and semantic alignment (Lyu et al., 2024; Thakur et al., 2024). To address these limitations, we introduce a retrieval-free evaluation paradigm tailored to Korean, enabling scalable, language-specific benchmarking of long-context reasoning without relying on retrievers.

3 Ko-LongRAG: A Korean Long-Context RAG Benchmark

We propose **Ko-LongRAG**, a high-quality **Ko**rean **Long**-context **RAG** benchmark designed to enable rigorous evaluation of long-context retrieval-based reasoning in Korean. We describe the benchmark creation process in Section §3.1, unanswerable cases—one of the key features of Ko-LongRAG—in Section §3.2, and benchmark statistics in Section §3.3. The detailed benchmark construction, including the prompts used, can be found in Appendix B.

3.1 Benchmark Construction

3.1.1 Document Clustering

Existing retrieval-based RAG methods struggle in low-resource languages due to limited structured knowledge bases and retrieval infrastructures. To address this, we introduce a novel Specialized Content Knowledge (SCK)-based document clustering and question-answer generation methodology, enabling high-quality long-context evaluation without an explicit retrieval step.

SCK encompasses a document's domain, central concepts, and key terms, facilitating structured content classification. We utilize the *DeepSeek v2.5* (DeepSeek-AI, 2024) model to extract SCK from the Korean Wiki Raw Corpus in JSON format. Based on the extracted SCK, we perform document clustering to categorize documents into structured domains. The distribution of each domain is shown in Figure 3, with a balanced representation across categories.

3.1.2 Long-context QA Generation

To generate meaningful question-answer pairs, we leverage *GPT-4o* (OpenAI, 2024a) model to create diverse and contextually rich questions (Abdullin et al., 2024). Using categorized documents, we generate two types of QA tasks: single-document QA, where questions are derived from individual documents requiring precise extraction, and multi-document QA, which involves reasoning across multiple documents within a cluster. Multi-document QA consists of comparison questions, which require factual comparisons across documents, and bridge questions, which demand logical inference by linking information from multiple sources. To construct long-context QA pairs, Ko-LongRAG clusters documents from the same domain as the source document, ensuring contextual consistency while maintaining diversity.

3.1.3 Manual Quality Assurance

To ensure the reliability of the dataset, Ko-LongRAG implements a manual review-based quality assurance process. Human annotators evaluate the generated QA pairs using a predefined checklist, and only those that fully meet the checklist criteria are retained in the final dataset. Details on the checklist, inter-annotator agreement (IAA), and the overall human annotation process are provided in Appendix B.4.

3.2 Incorporating Unanswerable Cases

A key feature of Ko-LongRAG is the incorporation of "unanswerable" cases, allowing for a robust evaluation of models' ability to handle uncertainty and retrieval failures. When an LLM encounters a document lacking sufficient information to answer a given query, it should explicitly indicate that an answer cannot be provided based on the given context. We systematically generate unanswerable cases by pairing documents with irrelevant questions to enforce this. We intentionally avoid hard-negative mining, as the clustered documents in Ko-LongRAG are already topically coherent, which may cause ambiguous overlaps between answerable and unanswerable cases. This design choice prioritizes clarity and ensures that unanswerable instances remain unambiguous, enabling more faithful evaluation of model robustness in clearly unsupported scenarios.

3.3 Benchmark Statistics

Both single-document QA and multi-document QA consist of 300 questions. For single-document QA, the average context length is 2,915 tokens, whereas multi-document QA extends to 14,092 tokens on average[2]. Additionally, 16.6% of the dataset consists of unanswerable questions. The dataset statistics are summarized in the Table 3 in Appendix B.5.

4 Experiments

4.1 Experimental Settings

We evaluate open-sourced and proprietary models on Ko-LongRAG to assess their long-context RAG performance in Korean. The open-sourced models in our experiments include multilingual models such as *Qwen 2.5* (Yang et al., 2024), *C4AI Command R* (Cohere For AI, 2024), *LLaMA 3* (Dubey et al., 2024), *Gemma 2* (Team et al., 2024), and *Phi 3* (Abdin et al., 2024), as well as Korean-specialized models like *EXAONE 3.5* (An et al., 2024), *SOLAR 10.7B* (Kim et al., 2023), and *LLaMa 3 Motif* (Moreh, 2024). Additionally, we conduct experiments with seven proprietary models (Achiam et al., 2023; Anthropic, 2024). The evaluation follows an LLM-as-a-Judge framework (Zheng et al., 2024), where *GPT-4o-2024-08-06* is used as the primary judge. Detailed experimental settings, including input and evaluation prompts, are provided in Appendix C.

4.2 Results

Overall Results Table 1 summarizes the overall performance of the evaluated models. Among open-sourced models, the *EXAONE 3.5* series achieves the highest performance across all parameter sizes, while among proprietary models, *o1-2024-12-17* records the best average score. In addition, among Korean-specialized open-sourced models, *EXAONE 3.5* consistently outperforms the others. Multi-document QA, which demands more complex reasoning, generally results in lower scores than single-document QA, highlighting its increased difficulty. Nevertheless, models that excel in single-document QA tend to retain their advantage in multi-document QA, suggesting that strong retrieval and comprehension skills carry over to multi-document reasoning.

Answerability Table 1 also presents results for unanswerable cases, evaluating how well models recognize the absence of an answer in the given document. Proprietary models generally perform well, whereas most open-sourced models struggle, with the exception of *EXAONE 3.5* and *Qwen 2.5*,

[2] We use OpenAI's tiktoken tokenizer for tokenization (`https://github.com/openai/tiktoken`).

Models	Single-doc QA			Multi-doc QA			Average
	Answerable	Unanswerable	Total	Answerable	Unanswerable	Total	
Open-sourced model (≥20B)							
EXAONE 3.5 32B	**92.4**	**100.0**	**93.7**	**72.8**	**98.0**	**77.0**	**85.3**
Qwen 2.5 32B	90.0	98.0	91.3	48.4	92.0	55.7	73.5
C4AI Command R 32B	85.6	66.0	82.3	62.4	62.0	62.3	72.3
Gemma 2 27B†	49.2	74.0	53.3	27.6	86.0	37.3	45.3
Yi 1.5 34B*	35.6	30.0	34.7	36.3	**98.0**	46.6	40.7
LLaMa-3-Motif 102B	34.8	92.0	44.3	12.8	86.0	25.0	34.7
Open-sourced model (∼10B)							
EXAONE 3.5 7.8B	68.4	**100.0**	73.7	**64.0**	**98.0**	**69.7**	**71.7**
LLaMa 3.1 8B	**78.0**	76.0	**77.7**	56.8	28.0	52.0	64.8
Qwen 2.5 7B	61.2	98.0	67.3	33.2	94.0	43.3	55.3
Gemma 2 9B†	30.4	**100.0**	42.0	26.4	90.0	37.0	39.5
Solar 10.7B‡	17.2	94.0	30.0	9.2	84.0	21.7	25.9
Phi 3 small (7B)	8.0	14.0	9.0	4.8	14.0	6.3	7.7
Open-sourced model (∼2B)							
EXAONE 3.5 2.4B	**80.8**	**100.0**	**84.0**	**61.6**	84.0	**65.3**	**74.7**
Qwen 2.5 3B	56.4	98.0	63.3	2.4	**94.0**	17.7	40.5
LLaMa 3.2 3B	48.8	12.0	42.7	40.0	16.0	36.0	39.3
Qwen 2.5 1.5B	22.0	96.0	34.3	21.6	92.0	33.3	33.8
Gemma 2 2B†	16.0	76.0	26.0	21.2	88.0	32.3	29.2
Proprietary model							
o1-2024-12-17	93.6	**100.0**	94.7	**88.0**	**100.0**	90.0	**92.3**
o1-mini-2024-09-12	87.2	**100.0**	89.3	85.2	**100.0**	87.7	88.5
GPT-4-turbo	90.4	**100.0**	92.0	76.0	96.0	79.3	85.7
GPT-4o-2024-11-20	**95.6**	**100.0**	**96.3**	68.0	**100.0**	73.3	84.8
GPT-4o-2024-08-06	95.2	**100.0**	96.0	63.2	**100.0**	69.3	82.7
Claude-3.5-Sonnet	78.4	**100.0**	82.0	73.7	**100.0**	78.1	80.1
GPT-4o-mini-2024-07-18	84.4	**100.0**	87.0	53.6	98.0	61.0	74.0

Table 1: Comparison results of language models on Ko-LongRAG benchmarks. The benchmark includes an "Unanswerable" case, where models must respond as "Unanswerable" if the answer is not in the context. **Bold** scores indicate the best performance, and underlined scores mean the second best. Context lengths: ‡ = 4k, † = 8k, * = 16k.

Models	Judges			Variance
	GPT-4o	o1-mini	Human	
Model-wise Results				
EXAONE 3.5 32B	85.3	85.1	85.2	**.0067**
QWEN 2.5 32B	73.5	73.2	73.0	**.0422**
C4AI Command R 32B	72.3	72.1	71.9	**.0267**
EXAONE 3.5 32B Intra-variance				
Repeat 3 ($n = 3$)	.0267	.1156	-	-
Repeat 5 ($n = 5$)	.0416	.0736	-	-
Repeat 7 ($n = 7$)	.1269	.0996	-	-

Table 2: Evaluation robustness analysis across different judge models and multiple repetitions. These results support the fact that Ko-LongRAG is a benchmark capable of robust evaluation.

which achieve near-perfect scores. This highlights their strong faithfulness in distinguishing unanswerable cases, as their responses closely align with the given context.

Separability Analyzing performance across different model sizes, we observe a positive correlation between parameter count and performance among open-sourced models. Specifically, increasing model size consistently improves results for *EXAONE 3.5*, *Qwen 2.5*, and *LLaMA 3* series models, except *EXAONE 3.5 2.4B*. This suggests that Ko-LongRAG scales with general language proficiency effectively, confirming its well-balanced difficulty distribution.

4.3 Evaluation Robustness

We conduct additional analysis to examine whether the experimental results remain consistent across various judge models and multiple experiment repetitions. Table 2 presents the robustness of the judge prompt. *GPT-4o*, *o1-mini*, and human evaluations exhibit consistently low variance close to 0, indicating stable reliability. Increasing the repeat count of *GPT-4o* and *o1-mini* judges for *EXAONE 3.5 32B* also results in minimal variance, further confirming the reliability of the evaluation.

5 Conclusion

We introduce Ko-LongRAG, the first Korean long-context RAG benchmark, addressing the lack of evaluation frameworks for non-English languages. Ko-LongRAG employs a retrieval-free approach using SCK to generate high-quality question-answer pairs. Through Ko-LongRAG, we evaluate the Korean long-context RAG performance of various LLMs, setting a new standard for Korean long-context RAG evaluation.

Limitations

Language Scope. Ko-LongRAG proposes a retrieval-free RAG benchmarking methodology tailored to low-resource languages, with its initial implementation targeting Korean. Although this demonstrates the feasibility of the approach within a single linguistic context, its broader applicability across diverse languages remains to be established. Extending the methodology to typologically distinct languages facilitates more comprehensive evaluation of multilingual RAG systems and supports the development of equitable, language-inclusive benchmarks.

Potential Distributional Bias. A widely adopted practice in benchmark construction involves generating QA pairs with high-performing language models and validating them manually to ensure quality. Following this approach, Ko-LongRAG employs GPT-4o, a strong proprietary model, for initial QA generation. While effective, this setup may introduce distributional bias that favors GPT-4o during evaluation. To address this concern, we validate all QA pairs manually and conduct comparative experiments using QA generated by two alternative proprietary models: Claude-3.5 Sonnet and Gemini-2.5 Flash (DeepMind, 2025). As detailed in Appendix C.4, the overall ranking trends remain consistent across QA sources, indicating that potential distributional bias has only minimal impact among proprietary models. Crucially, we observe that comparisons among open-source models remain stable regardless of the QA generation source, confirming that Ko-LongRAG provides a high-quality, bias-resilient dataset for reliable evaluation.

Ethics Statement

In our benchmark setup, we used publicly available datasets for their intended purposes. Furthermore, our evaluations with LLMs were conducted through their official websites, adhering to proper authorization protocols. All models utilized in our experiments were obtained from publicly accessible sources, including websites and GitHub repositories, in accordance with open science principles. Additionally, while drafting this paper, we leveraged an AI assistant to assist with sentence-level drafting and refinement. In addition, during the manual quality check process, it was confirmed that there is no potential risk in the benchmark.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. Synthetic dialogue dataset generation using llm agents. *arXiv preprint arXiv:2401.17461*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, et al. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv e-prints*, pages arXiv–2412.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. *arXiv preprint arXiv:2407.01463*.

Cohere For AI. 2024. c4ai-command-r-08-2024.

Google DeepMind. 2025. Gemini 2.5 pro.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

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? *arXiv preprint arXiv:2404.06654*.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*.

Moreh. 2024. Llama-3-motif-102b-instruct. Accessed: Feb. 16, 2025.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Introducing openai o1.

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

Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, and Amin Ahmad. 2024. Mirage-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems. *arXiv preprint arXiv:2410.13716*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Lei Zhang, Yunshui Li, Ziqiang Liu, Junhao Liu, Longze Chen, Run Luo, Min Yang, et al. 2023. Marathon: A race through the realm of long context with large language models. *arXiv preprint arXiv:2312.09542*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. 2025. Sglang: Efficient execution of structured language model programs. *Advances in Neural Information Processing Systems*, 37:62557–62583.

A Appendix: Ko-LongRAG Examples

Figure 7 presents examples from Ko-LongRAG, including both single-document QA and multi-document QA. For unanswerable cases, the reference answer is *"주어진 문서 내에서 답할 수 있는 정보가 충분하지 않습니다."* (*"The provided document does not contain sufficient information to answer this question"*). The context includes only the supporting documents from which an answer can be extracted. In multi-document QA, questions belong to the *comparison* type, requiring logical reasoning after comparing two documents. An example of such a question is: *"Which award offers a higher prize, the Asturias Award or the Miguel de Cervantes Prize"*?

B Appendix: Ko-LongRAG Details

This section introduces the benchmark details, including the prompts used for the LLMs employed in constructing Ko-LongRAG.

B.1 SCK Extraction Prompt

SCK Extraction Prompt

Document:

```
[begin]
{document}
[end]
```

Please recommend the key knowledge (e.g., Domain, Main Concept, Key Word) that should be considered the most important in the given document.
Your response must strictly follow this JSON format:

```
{"Domain": str, "Main Concept": str,
"Key Word": str}.
```

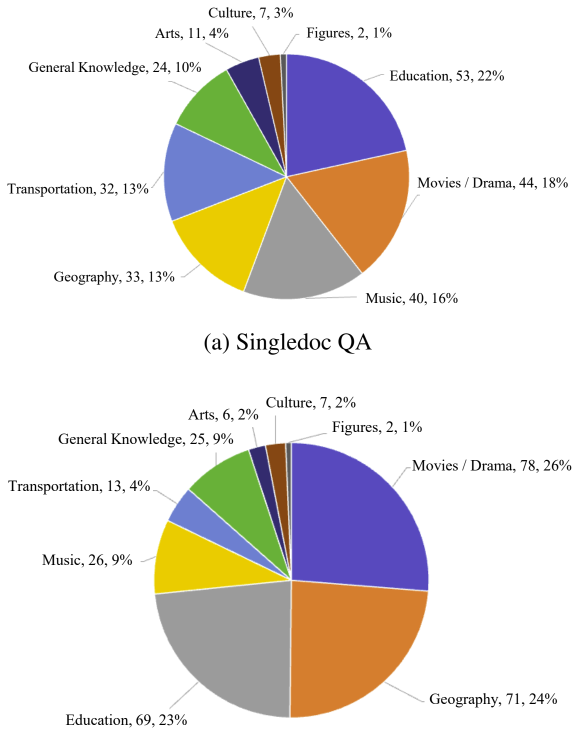
Please respond in Korean.

Figure 2: SCK Extraction Prompt.

To extract Specialized Content Knowledge (SCK) from documents, we employ a structured prompting approach using an LLM-based extraction method. As described in the main text, SCK consists of a document's domain, central concepts, and key terms, facilitating structured content classification. Figure 2 presents the prompt used for extracting SCK from the Korean Wiki Raw Corpus using the DeepSeek v2.5 model (DeepSeek-AI, 2024).

The prompt instructs the model to identify and extract the most critical knowledge elements from a given document while ensuring that the output follows a predefined JSON format. This structured format allows for systematic document clustering and domain categorization. Additionally, to maintain consistency and alignment with the dataset's language setting, the model is explicitly directed to respond in Korean. By leveraging this extraction method, we ensure that the generated SCK effectively captures the essential content structure of each document, providing a well-defined basis for domain classification and long-context evaluation.

B.2 Domain Distribution



(a) Singledoc QA

(b) Multidoc QA

Figure 3: Domain distribution.

Figure 3 illustrates the distribution of domains in Ko-LongRAG, ensuring a balanced representation across various fields. As described in the main text, documents are categorized into structured domains based on Specialized Content Knowledge (SCK), which includes domain, central concepts, and key terms. This classification is essential for maintaining content diversity and enabling reliable long-context evaluation.

In Figure 3 (a), the domain distribution for single-document QA is presented, while Figure 3 (b) shows the domain distribution for multi-document

QA. The dataset is designed to ensure that questions span multiple knowledge domains, allowing for a comprehensive evaluation of retrieval and reasoning capabilities. The balanced allocation of domains across both tasks supports robust generalization and minimizes potential dataset biases.

B.3 QA-Pair Generation Prompt

Document Pair Selection Prompt

Documents:

```
[begin]
{documents}
[end]
```

You are given a set of documents from the same domain.
Each document contains metadata including its **Main Concept** and **Keywords**.

Your task is to select a pair of documents that are suitable for creating a multi-document question.
There are two possible types:
- **Comparison-type:** comparing or contrasting two documents that cover similar concepts from different perspectives.
- **Bridge-type:** reasoning across two documents that are logically connected, where one builds upon or complements the other.

Please follow these instructions:

1. Read all documents and their metadata carefully.
2. Use the *Main Concept* and *Keywords* to analyze conceptual relationships.
3. Select two documents that are either comparison-type or bridge-type.
4. Briefly explain why you chose this pair based on their *Main Concepts*.

Figure 4: Document Pair Selection Prompt for Multi-document QA.

Ko-LongRAG employs an LLM-based approach using *GPT-4o* to create diverse and contextually rich questions. As described in the main text, two types of QA tasks are constructed: single-document QA, which requires precise extraction from individual documents, and multi-document QA, which involves reasoning across multiple documents within a thematically clustered set.

Before generating a multi-document QA pair, we prompt the LLM to select a document pair from the same domain cluster based on conceptual relationships inferred from the *Main Concepts* and *Keywords* within SCK. The prompt used for this step is shown in Figure 4. This selection step enables targeted generation of multi-hop questions while maintaining domain and conceptual coherence.

QA-Pair Generation Prompt

Single-doc QA:

You are a new question-answer pair maker.
Make the question more simpler.
But make them impossible to solve without reading the context carefully.
Questions should be short-answer questions.
Check the given context and existing problems and create new ones and corresponding answer.

Multi-doc QA:

You are a new question-answer pair maker.
Given two contexts, you'll need to create two types of questions.
Make them impossible to solve without reading the context carefully.
Questions must be short-answer format and Korean.

Type 1: Comparison question

Usually require contrasting two entities.
Example: Were Scott Derrickson and Ed Wood of the same nationality?

Type 2: Bridge question

Can be answered by following a connecting logic.
Example: Tysons Galleria is located in what county?

Figure 5: QA-Pair Generation Prompt.

Figure 5 presents the prompts used for question generation. For single-document QA, the prompt instructs the model to generate questions that require careful reading of the context while maintaining simplicity. For multi-document QA, the model is guided to create two distinct question types: (1) comparison questions, which contrast information across two documents, and (2) bridge questions, which require logical inference by linking information from multiple sources. In this case, the questions are generated based on document pairs selected through the procedure illustrated in Figure 4. These structured prompts ensure that the generated QA pairs comprehensively evaluate retrieval and reasoning capabilities in a long-context setting.

Human Annotation Checklist

1. Question Verification

Clarity: Is the question free from ambiguity or vague expressions?
Contextual Extractability: Can the question be answered solely based on the given context?
Grammar and Expression Errors: Does the question contain grammatical mistakes or unnatural expressions?
Multi-document QA: Does the question require referencing multiple contexts to be answered?

2. Answer Verification

Accuracy: Is the answer correct and appropriate for the given question?
Context-based Evaluation: Can the answer be fully derived from the entire context? (If only partially derivable, it is considered an incomplete question.)
Length Appropriateness: Is the answer excessively long, potentially affecting readability or evaluation?

3. Context Verification

Redundancy Removal: Does the context contain duplicate sentences that make it unnecessarily complex?
Grammar and Formatting: Are there any grammatical errors or formatting issues (e.g., spacing, spelling) in the context?

Figure 6: Human Annotation Checklist.

B.4 Human Annotation Checklist

To ensure the reliability of the dataset, Ko-LongRAG applies a manual review process where annotators verify the quality of generated question-answer (QA) pairs using a predefined checklist. Only QA pairs that fully meet these criteria are included in the final dataset.

Figure 6 presents the checklist, which covers three main aspects: question verification, answer verification, and context verification. Question verification checks for clarity, grammatical correctness, and whether the question can be answered using the given context. For multi-document QA, it also ensures that the question requires reasoning across multiple documents. Answer verification assesses whether the provided answer is accurate and fully supported by the context while maintaining the appropriate length. Context verification ensures that the provided text is free from redundancy, grammatical errors, and formatting issues.

To further enhance consistency and quality, the co-authors, who are proficient in Korean, serve as annotators and participate in a manual review process using the predefined quality checklist. Over the course of approximately two weeks, the annotators evaluate each QA pair based on three main criteria in the checklist. Only instances unanimously agreed upon by all three annotators are included in the final dataset. As a result, the inter-annotator agreement (IAA), calculated using Fleiss' Kappa across all instances including those that are excluded, is 0.77, which corresponds to a *Substantial* level of agreement.

B.5 Ko-LongRAG Statistics

Category	QA Type	
	Single-document	Multi-document
Number of Questions	300	300
Context Length (tokens)	2,915	14,092
Answerability	250 answerable, 50 unanswerable	

Table 3: Ko-LongRAG Benchmark Statistics.

A detailed summary of benchmark statistics is presented in Table 3.

C Appendix: Evaluation Details

C.1 Experimental Settings

All model inference is conducted using the `sglang` inference engine (Zheng et al., 2025), with all prompts formulated in Korean to ensure proper Korean language processing. Open-sourced models are inferred using eight NVIDIA A100 GPUs, while proprietary models are evaluated using the default settings provided by their respective model APIs. Given that Ko-LongRAG has an average token length of 14k, we apply middle truncation for models with shorter maximum context lengths of 4k, 8k, or 16k—such as *SOLAR 10.7B*, *Gemma 2*, and *Yi-chat 34B*—to accommodate longer contexts.

The middle truncation (Bai et al., 2023) preserves the first and last segments of input while discarding the middle portion, ensuring that both introductory and concluding details remain accessible. This approach helps mitigate information loss while adhering to the context length limitations of the models.

C.2 Ko-LongRAG Prompt

We provide the prompt used for Ko-LongRAG evaluation in Figure 8. We designed this prompt to ensure that models generate responses strictly based on the given document.

At the end of the prompt, the following instruction is included: *"답변을 문서에서 찾을 수 없는 경우, '주어진 정보로 답할 수 없다'로 응답하세요"* *"(If the answer cannot be found in the document, respond with 'The provided information does not allow for an answer.'")*. This instruction ensures that models correctly handle unanswerable cases, where responses like *"The provided document does not contain sufficient information to answer this question."* are considered correct.

C.3 Ko-LongRAG LLM-as-a-Judge Prompt

The LLM-as-a-Judge prompt used for evaluating Ko-LongRAG benchmark performance is provided in Figure 9.

C.4 Bias Mitigation through Alternative QA Generation

Models	QA Generator		
	GPT-4o	Claude-3.5	Gemini-2.5
	Proprietary model		
o1	92.3	91.2	90.9
GPT-4o	84.8	85.6	83.4
Claude-3.5	80.1	81.4	79.7

Table 4: Bias mitigation results across QA generators. Model and generator names are abbreviated for clarity: **o1** = *o1-2024-12-17*; **GPT-4o** = *GPT-4o-2024-11-20*; **Claude-3.5** refers to *Claude-3.5 Sonnet*; **Gemini-2.5** refers to *Gemini-2.5 Flash*.

To evaluate whether our benchmark exhibits distributional bias in favor of GPT-4o due to its role in data construction, we conduct a supplementary experiment using QA pairs generated by two alternative proprietary models: Claude 3.5 Sonnet and Gemini 2.5 Flash. In the case of Claude-3.5 Sonnet, which may benefit from distributional alignment when used to generate data, we observe that GPT-4o still outperforms Claude-3.5 Sonnet, preserving the original ranking order. For Gemini-2.5 Flash, a strong frontier model used solely for QA generation and not included in the set of evaluated models, therefore free from distributional bias, we find that model rankings remain consistent, further confirming the robustness of the benchmark.

These findings suggest that while any frontier model used for QA generation may introduce slight bias in its favor, such effects do not significantly distort evaluation outcomes. In particular, model comparisons involving open-source LLMs remain stable and meaningful, reinforcing the benchmark's utility as a fair and reliable testing ground across diverse model families.

Ko-LongRAG Examples

[Single-doc QA Answerable Case]

Context:

...

Title: 박진우 (야구인)

Text: 박진우(朴晋佑, 1990년 2월 12일 ～)는 전 KBO 리그 NC 다이노스의 투수이자, 현 KBO 리그 SSG 랜더스의 스카우트이다.

...

2019년 시즌 : 선발과 불펜을 가리지 않고 활약했다. 시즌 140.2이닝 3점대 평균자책점, 92탈삼진, 9승 7패, 5홀드를 기록했다. 이동욱 감독은 '가장 MVP로 꼽고 싶은 선수'라며 칭찬했다.

...

Question: 박진우가 NC 다이노스에서 9승을 기록한 시즌은 언제인가요?
Answer: 2019년

[Single-doc QA Unanswerable Case]

Question: 인천남동소방서의 설립 연도는 무엇인가요?
Answer: 주어진 문서내에서 답할 수 있는 정보가 충분하지 않습니다.

[Multi-doc QA Answerable Case]

Context:

...

Title: 아스투리아스 공상

Text: 아스투리아스 공상은 스페인의 프린시페 데 아스투리아스 재단(Fundación Príncipe de Asturias)이 주관하는 상이다. 1980년 9월 24일 스페인의 왕세자에 해당하는 호칭인 아스투리아스 공이었던 펠리페 (Felipe, 펠리페 6세)에 의해 제정되었으며 1981년에 첫 시상식이 열렸다. 총 9개 부문 (예술 부문, 커뮤니케이션 · 인문주의 부문, 국제 협력 부문, 문학 부문, 사회과학 부문, 체육 부문, 기술 · 과학 연구 부문, 화합 부문, 아스투리아스 모범상 부문)으로 나누어 시상한다. 시상식은 아스투리아스 지방의 오비에도에서 열린다. 수상자는 주안 미로가 제작한 조각, 상금 50,000 유로를 받게 된다.

...

Title: 미겔 데 세르반테스 상

Text: 미겔 데 세르반테스 상(-賞,) 또는 세르반테스 상은 스페인 작가 미겔 데 세르반테스의 이름이 붙은 스페인어 작가에게 수여되는 문학상으로, 영연방의 맨 부커 상과 유사한 스페인어권의 상이다. 그러나 맨 부커 상과는 다르게 일생 동안의 문학적 성취를 평가해서 단 한 번만 수여하므로 스페인어권에서 그 권위는 노벨 문학상에 버금간다. 1976년 제정되었다. 스페인 문화부가 수여하며 상금은 12만 5천 유로이다.

...

Question: 아스투리아스 공상과 미겔 데 세르반테스 상 중 상금이 더 많은 것은 무엇인가요?
Answer: 미겔 데 세르반테스 상

[Multi-doc QA Unanswerable Case]

Question: 넬슨 록펠러와 노아 사이러스는 둘 다 정치 경력을 가지고 있었나요?
Answer: 주어진 문서내에서 답할 수 있는 정보가 충분하지 않습니다.

Figure 7: Examples of Ko-LongRAG.

Ko-LongRAG Prompt

[Single-doc QA Prompt]

System: 당신은 도움이 되는 어시스턴트입니다.
User:
다음 문서를 살펴보고, 질문에 대한 답을 추출하세요.
질문에 대한 답만 생성하세요. 답변은 매우 간결해야 합니다.
답변을 문서에서 찾을 수 없는 경우,
’주어진 정보로 답할 수 없다’로 응답하세요.

문서는 “Title”에 따라 정렬된 Wikipedia 문단 목록이며 제목은 다음과 같습니다: {{*titles*}}.
각 위키피디아 문단은 ’Title’ 필드와 ’Text’ 필드를 포함합니다.
문서는 다음과 같습니다: {{*context*}}. 질문은 다음과 같습니다: {{*question*}}.

[Multi-doc QA Prompt]

System: 당신은 도움이 되는 어시스턴트입니다.
User:
다음 문서를 검토하고 질문에 답하세요.
문서는 Wikipedia 문단 목록이며 제목은 다음과 같습니다: {{*titles*}}.
질문에는 두 가지 유형이 있습니다:
예 또는 아니오로 답하거나 두 후보 중에서 선택해야 하는 비교 질문과, 단답형 형태의 일반 질문입니다.
문서는 다음과 같습니다: {{*context*}}.
문서에서 필요한 문단을 찾아 질문에 답하세요: {{*question*}}.
일반 질문의 경우 문단에서 정확한 단어를 찾아서 답변해야 합니다.
질문에 대한 답만 생성하고 다른 어떤 것도 생성하지 마세요.
답변을 문서에서 찾을 수 없는 경우, ’주어진 정보로 답할 수 없다’로 응답하세요.

Figure 8: Prompt for evaluating Ko-LongRAG.

Ko-LongRAG LLM-as-a-Judge Prompt

System:

You are an expert evaluator of text answers in Korean.
Your task is to compare the content of two Korean answers, a long answer (`long_ans`) and a short answer (`short_ans`), with the provided correct answers (`Answer`), which may contain multiple correct options.
Both the long answer and the short answer need to be checked for correctness. The long and short answers do not need to match any of the answers in the `Answer` list word-for-word but must convey the same key meaning or idea.
If either the long or short answer matches any one of the correct answers in the `Answer` list, it should be considered correct.
Focus only on the accuracy of the content and ignore style, tone, or extra information unless it introduces inaccuracies.
For both the long and short answers, return only the evaluation result as a Python dictionary object, and ensure the output is formatted as valid Python code.

Here are two examples of how to evaluate answers:

Example 1:
`Question`: HP는 게임에서 무엇을 의미하나요?
`Answer`: ['체력', '생명력']
`long_ans`: HP는 ’생명력’ 또는 ’체력’을 의미하며, 게임에서 캐릭터의 생존력을 나타내는 지표입니다. HP가 줄어들면 캐릭터는 점점 약해지며, 0이 되면 게임에서 탈락하거나 패배할 수 있습니다.
`short_ans`: HP는 캐릭터의 체력입니다.
`Evaluation`: {'long_ans': 'correct', 'short_ans': 'correct'}

Example 2:
`Question`: 프랑스의 수도는 어디인가요?
`Answer`: ['파리']
`long_ans`: 프랑스의 수도는 파리로, 리옹의 오른쪽 아래에 위치하고, 문화와 예술의 중심지로 알려져 있습니다. 에펠탑, 루브르 박물관, 노트르담 대성당 등 유명한 관광지가 위치해 있습니다.
`short_ans`: 프랑스의 수도는 리옹입니다.
`Evaluation`: {'long_ans': 'correct', 'short_ans': 'incorrect'}

Now, proceed with your evaluation of the following question, answer, and responses, and return only the evaluation as a valid Python dictionary.
Ensure the response is a valid Python dictionary object without any additional text.

User:

Evaluate the following long and short answers based on the provided correct answer.
Your goal is to determine if the long and short answers are correct.
Return the evaluation result in the form of a Python dictionary: `{'long_ans': 'correct 'or 'incorrect ', 'short_ans': 'correct'or 'incorrect'}`.

`Question`: {{*question*}}
`Answer`: {{*answer*}}
`long_ans`: {{*long_ans*}}
`short_ans`: {{*short_ans*}}

Return only the evaluation in the form of a Python dictionary.
Do not include any explanation or additional comments.

Figure 9: LLM-as-a-judge prompt for evaluating Ko-LongRAG.