EXAONE 3.5:
Series of Large Language Models for Real-world Use Cases

LG AI Research[*]

Abstract

This technical report introduces the EXAONE 3.5 instruction-tuned language models, developed and released by LG AI Research. The EXAONE 3.5 language models are offered in three configurations: 32B, 7.8B, and 2.4B. These models feature several standout capabilities: 1) exceptional instruction following capabilities in real-world scenarios, achieving the highest scores across seven benchmarks, 2) outstanding long-context comprehension, attaining the top performance in four benchmarks, and 3) competitive results compared to state-of-the-art open models of similar sizes across nine general benchmarks. The EXAONE 3.5 language models are open to anyone for research purposes and can be downloaded from `https://huggingface.co/LGAI-EXAONE`. For commercial use, please reach out to the official contact point of LG AI Research: contact_us@lgresearch.ai.

1 Introduction

EXAONE 3.0 instruction-tuned large language model with 7.8B parameters [41] demonstrated strong bilingual capabilities in Korean and English with exceptional real-world performance and instruction-following proficiency. Since its release, we have received diverse feedback from both academic and industrial communities. For instance, academic researchers have emphasized the need for smaller models that can be trained and deployed on low-specification GPUs due to limited access to advanced computational infrastructure. The industry has expressed strong demand for larger models with enhanced performance that remain cost-effective, as well as smaller models suitable for on-device deployment. Additionally, with the increasing adoption of retrieval-augmented generation (RAG) techniques, which generate answers based on reference documents or web search results, there has been substantial demand for models capable of effectively handling longer contexts.

In this report, we present EXAONE 3.5 language models, a collection of instruction-tuned language models ranging from 2.4B to 32B parameters, developed to meet the diverse needs of users. EXAONE 3.5 language models include: 1) 2.4B model optimized for deployment on small or resource-constrained devices, 2) 7.8B model matching the size of its predecessor but offering improved performance, and 3) 32B model delivering exceptional performance. All models support long-context processing of up to 32K tokens. Each model demonstrates state-of-the-art performance in real-world use cases and long-context handling, while remaining competitive in general domains compared to recently released models of similar sizes.

With the release of the EXAONE 3.5 language models, we hope to support researchers to push the boundaries of generative AI and inspire the development of innovative applications that enhance human life. This is in line with the mission of LG AI Research: ADVANCING AI FOR A BETTER LIFE.

2 Model Training

This section describes the detailed information on model configurations and the methods used for pre-training and post-training phases, along with the dataset construction process for each training phase.

[*] The complete list of authors who contributed to this work can be found in Appendix A.

Model size	32B	7.8B	2.4B
d_model	5,120	4,096	2,560
Number of layers	64	32	30
Pre-normalization	True	True	True
Non-linearity	SwiGLU [44]	SwiGLU	SwiGLU
Feedforward dimension	27,392	14,336	7,168
Head type	GQA [3]	GQA	GQA
Number of heads	40	32	32
Number of KV heads	8	8	8
Head size	128	128	80
Max sequence length	32,768	32,768	32,768
RoPE theta [46]	1,000,000	1,000,000	1,000,000
Tokenizer	BBPE [51]	BBPE	BBPE
Vocab size	102,400	102,400	102,400
Tied word embedding	False	False	True

Table 1: Configurations of EXAONE 3.5 language models

2.1 Model Configurations

The EXAONE 3.5 language models are based on the latest decoder-only Transformer architecture, and detailed configurations are described in Table 1. These models are identical in structure to the EXAONE 3.0 7.8B model but mainly differ in their configurations related to sizes. Notably, the EXAONE 3.5 language models extend the maximum context length from 4,096 tokens in EXAONE 3.0 to 32,768 tokens by adopting the long-context fine-tuning [7]. All three models share the same vocabulary, which consists roughly of 50% Korean and 50% English.

2.2 Pre-training

The amount of pre-training corpus data and computational resources are shown in Table 2. The approach to data construction and model training consists of two stages: 1) we perform first-stage pre-training based on the large training corpus, which is collected and processed from as diverse sources as possible aimed to increase the performance on general domains. After that, 2) we collect more data for the domains that need to be strengthened through evaluations and conduct second-stage of pre-training. For instance, we focus on enhancing long-context understanding capabilities in the second-stage.

Model size	32B	7.8B	2.4B
Training tokens	6.5T	9T	6.5T
Amount of computation (FLOPs)	1.25×10^{24}	4.21×10^{23}	9.36×10^{22}

Table 2: The sizes of the training data corpus along with the amounts of computation to build EXAONE 3.5 language models

2.2.1 Context Length Extension

To extend the context length, we utilize the long-context fine-tuning technique [7]. To mitigate the catastrophic forgetting problem [32], where the model forgets what it learned during the first pre-training stage, a replay-based method [2] is applied. Specifically, during the second-stage pre-training, we reuse a portion of the data used in the first-stage. While documents exceeding the maximum context length is split into smaller chunks in the first-stage, the original corpus are trained without being divided into chunks in the second-stage to extend the models' context length.

2.2.2 Decontamination

By the nature of massively web-crawled corpus, test-set examples often appear in the training corpus [43, 55]. These contaminated examples are likely to harm generalization performance and confuse test metrics, thus presenting unfair evaluations to users. To prevent the contaminated examples undermine the generalization performance of EXAONE

3.5 language models, we rigorously apply a decontamination process for all targeted benchmark test data and remove contaminated examples from the training pipeline.

We borrow a simple yet powerful substring-level matching method [36] with stricter criteria. The entire decontamination process is described in Figure 4 in Appendix C. We first normalize all test-set examples by removing all other characters except alphabets and numbers, then we extract all unique substrings with sliding window size $S = 50$ and a stride of 1. To determine whether a training example is contaminated, we randomly sample $N = 10$ substrings from the normalized training example and check if they exist in the substring pools. Table 10 in Appendix C provides examples of documents found in web corpora considered as contaminated.

2.2.3 Training Cost

Considering the computational cost of pre-training a large language model (LLM), it is necessary to make the training efficient by achieving as much high performance as possible with limited resources. Table 3 compares the total amounts of computations required for pre-training between the EXAONE 3.5 32B language model and others of similar size. When we simply approximate the total amounts of computations as the product of the model size and the number of training tokens [19, 24], Qwen 2.5 32B, for example, requires 2.77 times more computations than EXAONE 3.5 32B. One of the noticeable characteristics of the EXAONE 3.5 language models is that they demonstrate high performance despite being trained at lower costs than the other baseline models (see Section 3).

Models	Model size	Training tokens	Amount of computation (ratio)
EXAONE 3.5	32B	6.5T	1.00
Qwen 2.5	32B	18T	2.77
Gemma 2	27B	13T	1.69
Yi 1.5	34B	3.6T	0.59

Table 3: Comparison of the total amounts of computations to build models. We approximate the amount of computations as the product of the model size and the number of training tokens. Although the EXAONE 3.5 32B model is behind in the computations compared to Qwen 2.5 and Gemma 2, it has shown competitive performances.

2.3 Post-training

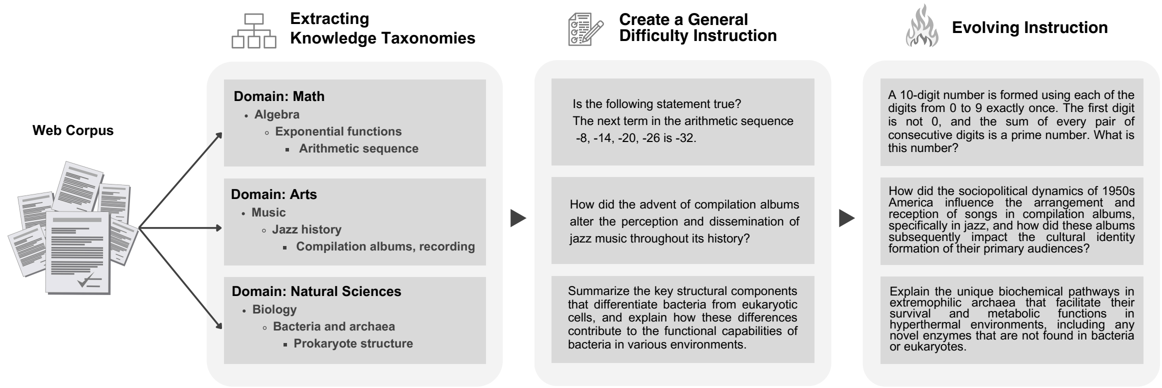
After pre-training, models go through further processes for strengthening their instruction-following capabilities and aligning with human preferences, which are well known as supervised fine-tuning (SFT) [53] and preference optimization.

2.3.1 Supervised Fine-tuning

To perform well on new or unseen instructions, a model needs to be trained on pairs of instruction-response datasets with varying difficulty from different domains. Hence, in order to build training data covering a wide range of fields, we extract core knowledge from 8M web corpora using a taxonomic system, as shown in Figure 1. We then generate an instruction-tuning dataset based on the extracted knowledge taxonomy. Finally, leveraging an instruction evolution method, which stems from the method proposed in [58], we diversify the complexity levels so that instructions with various complexities and difficulties can be produced.

2.3.2 Preference Optimization

Direct alignment algorithms (DAAs) [38], such as DPO [39] and SimPO [33], are used to train models after supervised fine-tuning to align models with human preferences. We create preference data for training using synthetic data and pre-collected data. For response generation, we sample N responses from multiple models for the prompt x drawn from the preference data and select the best response as y_w and the worst response as y_l based on the scores of a reward model to create a preference data, $\{x, y_w, y_l\}$. To validate preference data, we use an additional reward model to calculate agreement based on the rankings of the two reward models and filter out data with agreement below the threshold. Our preference optimization comprises multiple stages to sequentially train models M_1 and M_2 through DAAs, where M_0 is initialized from the SFT model. The staged pipeline enables us to mitigate over-optimization [38] that may occur during the DAAs' training process. Figure 2 shows a schematic diagram for constructing our preference dataset and training process.



Is the following statement true? The next term in the arithmetic sequence -8, -14, -20, -26 is -32.

How did the advent of compilation albums alter the perception and dissemination of jazz music throughout its history?

Summarize the key structural components that differentiate bacteria from eukaryotic cells, and explain how these differences contribute to the functional capabilities of bacteria in various environments.

A 10-digit number is formed using each of the digits from 0 to 9 exactly once. The first digit is not 0, and the sum of every pair of consecutive digits is a prime number. What is this number?

How did the sociopolitical dynamics of 1950s America influence the arrangement and reception of songs in compilation albums, specifically in jazz, and how did these albums subsequently impact the cultural identity formation of their primary audiences?

Explain the unique biochemical pathways in extremophilic archaea that facilitate their survival and metabolic functions in hyperthermal environments, including any novel enzymes that are not found in bacteria or eukaryotes.

Figure 1: A procedure of instruction-tuning data construction. First, we extract the core knowledge from large-volume web corpora and classify it within the taxonomy we defined in advance. Next, instruction-tuning data is generated based on the knowledge. To construct additional training data that is more complex, we leverage an instruction-evolving method [58] that lets the final dataset cover various fields with varying levels of difficulty.

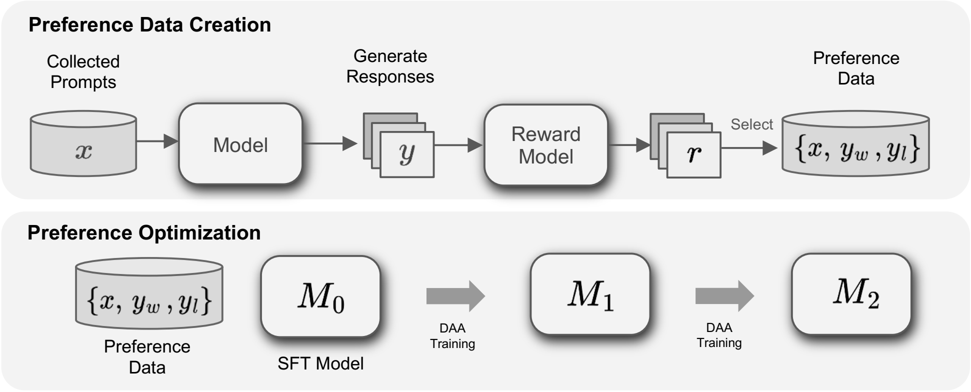


Figure 2: Overview of the preference optimization pipeline. (Top) Preference Data Creation: It shows the process of constructing preference data $\{x, y_w, y_l\}$ by scoring the responses y generated from a model for the prompt x using a reward model. (Bottom) Preference Optimization: Sequential training process where M_0 initialized from the SFT model is trained through DAA to obtain M_1 and M_2.

2.4 Data Compliance

Developing AI models requires a large amount of data, and the acquisition and utilization of this data can lead to various legal issues, such as copyright infringement, intellectual property infringement, and personal information protection violations. To minimize these risks, LG AI Research conducts AI Compliance reviews throughout the entire process of data collection, AI model training, and information provision. For more detailed information, please refer to the EXAONE 3.0 Technical Report [41] and the LG AI Ethics Principles [28].

3 Evaluation

This section presents the evaluation settings and results of EXAONE 3.5 language models on various benchmark datasets. We select recently released open-sourced language models for baselines of our models to compare our performances on the benchmarks. All baselines and their detailed information are described in Appendix D.1.

3.1 Benchmarks

Considering the diverse nature of user intents, it is crucial for an instruction-tuned model to generate a response aligned to the user's query, whatever it is. To evaluate our models in comprehensive and various scenarios, we select over a dozen evaluating benchmarks along with a few in-house benchmarks. Table 4 summarizes all benchmarks, which can be grouped into three categories:

- **Real-world Use Cases** (Section 3.3): the benchmarks requiring the ability to understand and perform diverse user instructions.
- **Long Context** (Section 3.4): the benchmarks evaluating the ability to understand the long context.
- **General Domain** (Section 3.5): the benchmarks embracing general domain abilities that LLMs are expected to have. Specifically, this category includes benchmarks for measuring the ability to solve mathematical problems, the ability to write source codes, and the parametric knowledge embedded in an LLM.

Category	Benchmark	Lang	Evaluation Settings	Metric
Real-world Use Cases	MT-Bench [59]	EN	LLM-as-a-judge (judge: *gpt-4o-2024-08-06*)[1]	LLM score
	LiveBench [54] (v2024-08-31)	EN	Ground-truth match	Accuracy
	Arena-Hard-v0.1 [29]	EN	LLM-as-a-judge (judge: *gpt-4-1106-preview*)	Win rate
	AlpacaEval 2.0 LC [12]	EN	LLM-as-a-judge (judge: *gpt-4-1106-preview*)	Win rate
	IFEval[61]	EN	Prompt-level / strict accuracy	Accuracy
	KoMT-Bench [42]	KO	LLM-as-a-judge (judge: *gpt-4o-2024-08-06*)	LLM score
	LogicKor [37]	KO	LLM-as-a-judge (judge: *gpt-4-1106-preview*)	LLM score
Long Context	Needle-In-A-Haystack [23]	EN/KO	Ground-truth match	Accuracy
	LongBench [5]	EN	Ground-truth match	F1, Rouge
	LongRAG [21] (extended)	EN	LLM-as-a-judge (judge: *gpt-4o-2024-08-06*)	LLM score
	Ko-LongRAG (In-house)	KO	LLM-as-a-judge (judge: *gpt-4o-2024-08-06*)	LLM score
	Ko-WebRAG (In-house)	KO	LLM-as-a-judge (judge: *gpt-4o-2024-08-06*)	LLM score
General Domain	GSM8K [9]	EN	0-shot / CoT	Accuracy
	MATH [17, 27]	EN	0-shot / CoT	Accuracy
	HumanEval [6]	EN	0-shot	pass@1
	MBPP [4]	EN	0-shot (Evalplus base)[2]	pass@1
	GPQA [40]	EN	0-shot / CoT	Accuracy
	ARC-C [8]	EN	0-shot	Accuracy
	BBH [47]	EN	0-shot / CoT	Accuracy
	MMLU [16]	EN	0-shot / CoT	Accuracy
	KMMLU [45]	KO	0-shot / CoT	Accuracy

Table 4: The benchmarks used to evaluate the performance of EXAONE 3.5 language models along with their target languages, evaluation settings, and the metrics. LONGRAG is extended from the original, and KO-LONGRAG and KO-WEBRAG are in-house benchmarks (see Section 3.4).

3.2 Overall Performance

The results of overall performance against three categories are presented in Table 5. Our EXAONE 3.5 language models, with sizes 32B and 7.8B, perform best in Real-world Use Cases and Long Context categories compared to baseline models while showing competitive results in the General Domain category. Our smallest model, EXAONE 3.5 2.4B, outperforms baselines with similar sizes in all three categories, demonstrating strong performance. Surprisingly, our 2.4B model, despite its small size, has shown better performance compared to baselines even with a larger size (< 9B) except for Qwen 2.5 7B in General Domain. Considering the recent surge in demand for smaller large language models (sLLM) [52], we believe that our EXAONE 3.5 2.4B model is well-positioned to be highly competitive in both academic and industrial use.

In the following sections, we elaborate on detailed evaluation settings and the results for each category.

3.3 Real-world Use Cases

For the Real-world Use Cases category, we have compiled seven benchmarks that represent real-world queries users might submit to a chatbot model. In MT-BENCH, KOMT-BENCH, and LOGICKOR, models' responses consisting of multi-turns are evaluated by a judge model. For ARENA-HARD and ALPACAEVAL, responses of a target language

[1] The separability of the original GPT-4 judge results is notably low, prompting the adoption of *gpt-4o-2024-08-06* as judge.

[2] We choose the MBPP base from EvalPlus [31], which is a subset of the original and consists of refined, high-quality problems.

Models	Real-world Use Cases	Long Context	General Domain
EXAONE 3.5 32B	**74.3**	**71.1**	74.8
Qwen 2.5 32B [49]	69.8	66.9	**78.7**
C4AI Command R 32B [10]	46.0	63.4	56.8
Gemma 2 27B [48]	64.2	-	68.7
Yi 1.5 34B [2]	46.9	-	53.9
EXAONE 3.5 7.8B	**70.7**	**66.6**	70.2
Qwen 2.5 7B [49]	52.7	56.1	**71.0**
Llama 3.1 8B [15]	48.6	58.8	62.4
Gemma 2 9B [48]	57.9	-	62.9
Phi 3 small (7B) [1]	41.7	33.4	63.2
EXAONE 3.5 2.4B	**61.1**	**63.4**	**63.3**
Qwen 2.5 3B [49]	44.5	40.7	62.1
Qwen 2.5 1.5B [49]	30.1	34.5	47.9
Llama 3.2 3B [34]	36.7	44.2	54.9
Gemma 2 2B [48]	41.7	-	42.2

Table 5: Overall comparison results of EXAONE 3.5 language models with similar-sized baseline language models. Here, a dash (-) indicates the model does not support context lengths longer than 16K. **Bold** scores indicate the best performance, and underlined scores mean the second best. The detailed information for each baseline is described in Appendix D.1.

model are compared with those of a reference model (*gpt-4-0314* and *gpt-4-1106-preview*, respectively) by a judge model, recording the win rate. LiveBench (ver. 2024-08-31) and IFEval (prompt-strict) assess how well the models' responses align with user instructions by matching them to the ground-truth responses.

Models	MT-Bench	LiveBench	Arena-Hard	AlpacaEval	IFEval	KoMT-Bench	LogicKor	Average
EXAONE 3.5 32B	**8.51**	43.0	**78.6**	**60.6**	**81.7**	**8.05**	**9.06**	**74.3**
Qwen 2.5 32B	8.49	**50.6**	67.0	41.0	78.7	7.75	8.89	69.8
C4AI Command R 32B	7.38	29.7	17.0	25.9	26.1	6.72	8.24	46.0
Gemma 2 27B	8.28	40.0	57.5	52.2	59.7	7.19	8.56	64.2
Yi 1.5 34B	7.64	26.2	23.1	34.8	55.5	4.88	6.33	46.9
EXAONE 3.5 7.8B	**8.29**	**39.8**	**68.7**	**54.2**	**78.9**	**7.96**	**9.08**	**70.7**
Qwen 2.5 7B	6.48	35.6	48.9	31.7	72.5	5.19	6.38	52.7
Llama 3.1 8B	7.59	28.3	27.7	25.7	74.5	4.85	5.99	48.6
Gemma 2 9B	7.64	32.1	43.6	47.3	54.7	7.10	8.05	57.9
Phi 3 small (7B)	7.63	27.9	26.8	29.2	59.5	3.22	3.99	41.7
EXAONE 3.5 2.4B	**7.81**	**33.0**	**48.2**	**37.1**	**73.6**	**7.24**	**8.51**	**61.1**
Qwen 2.5 3B	7.21	25.7	26.4	17.4	60.8	5.68	5.21	44.5
Qwen 2.5 1.5B	5.72	19.2	10.6	8.4	40.7	3.87	3.60	30.1
Llama 3.2 3B	6.94	24.0	14.2	18.7	70.1	3.16	2.86	36.7
Gemma 2 2B	7.20	20.0	19.1	29.1	50.5	4.83	5.29	41.7

Table 6: Performance comparison results of EXAONE 3.5 language models with similar-sized recently-released language models on seven benchmarks representing real-world use case scenarios. When calculating the macro average, the scores of MT-Bench, KoMT-Bench, and LogicKor are multiplied by 10 because they are scored out of 10 and the rest are scored out of 100. **Bold** scores indicate the best performance, and underlined scores mean the second best.

As presented in Table 6, our three models have shown the best performance against baselines of similar size in all benchmarks, except for the 32B model in LiveBench. Furthermore, by outperforming others in both English and Korean benchmarks, EXAONE 3.5 language models demonstrate their superior bilingual abilities.

3.4 Long Context

The ability to process and understand long contexts is increasingly important for modern LLMs, as it enables their application in more complex scenarios. To demonstrate EXAONE language model's long context performance, we

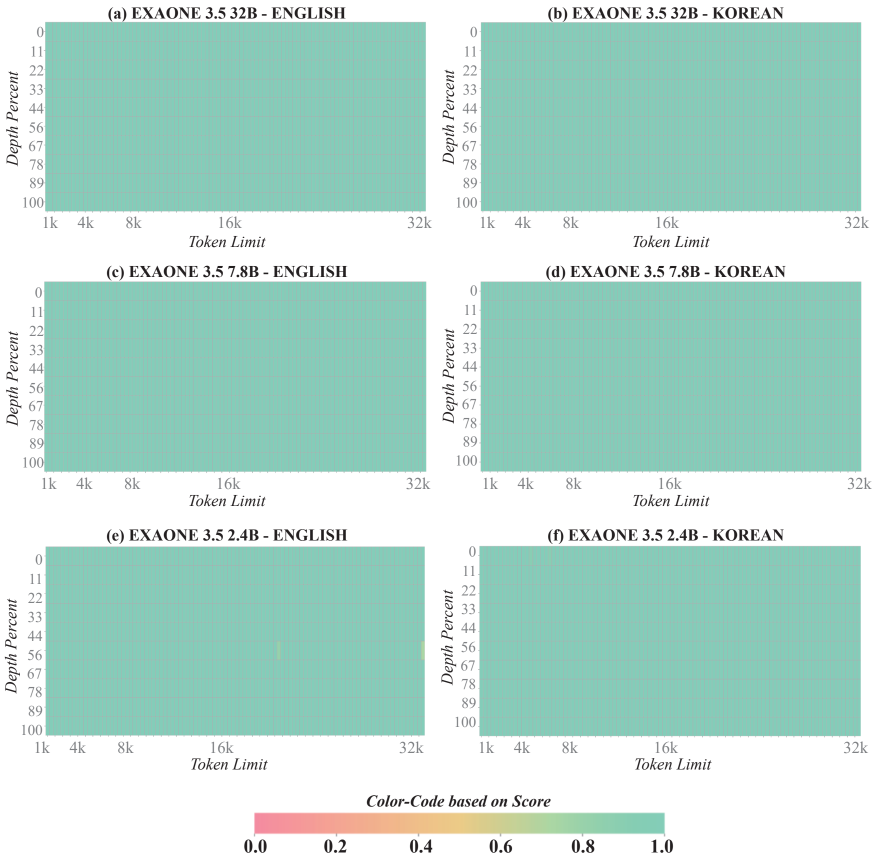


Figure 3: NIAH results of EXAONE 3.5 language models. The x-axis represents the token length of the input text, while the y-axis shows the relative position within the text, expressed as a percentage (0% corresponds to the beginning, and 100% to the end). The results are represented using a color-coded scheme: green indicates successful retrievals, and red represents unsuccessful ones. EXAONE 3.5 language models achieve near-perfect accuracy in retrieving information across various document depths and context lengths in English and Korean.

evaluate our models using benchmarks designed for a synthetic task with long context inputs, along with various retrieval-augmented generation (RAG) benchmarks.

3.4.1 Needle-in-a-Haystack

Needle-in-a-Haystack (NIAH) [23] serves as a benchmark to assess how effectively models can locate and retrieve information hidden at random locations within long documents. We comprehensively evaluate our models' ability to process and retrieve information from long contexts, up to 32K tokens. Furthermore, we extend NIAH to Korean and employ it to evaluate our models' long context processing ability across both English and Korean contexts.

Figure 3 demonstrates that our models achieve near-perfect accuracy in retrieving targeted information across all tested document depths and context lengths in both English and Korean. These results highlight their robust long context processing capabilities, particularly in tasks demanding precise information retrieval and complex reasoning.

3.4.2 Long Context Understanding

To assess long context understanding capabilities, we evaluate our models using benchmarks including LongBench [5] and LongRAG [21]. We expand unanswerable cases in LongRAG to make it more challenging. We also build Ko-LongRAG, the Korean counterpart to LongRAG, to evaluate long context understanding in Korean. For more realistic RAG scenario, requiring answers to difficult questions using actual web-searched results, we constructed Ko-WebRAG benchmark. We refer readers to the Appendix D.2 for more details.

Models	LongBench	LongRAG	Ko-LongRAG	Ko-WebRAG	Average
EXAONE 3.5 32B	49.2	**67.6**	**85.3**	**82.3**	**71.1**
Qwen 2.5 32B	49.1	63.6	73.5	81.3	66.9
C4AI Command R 32B	**50.9**	55.3	72.3	75.0	63.4
Gemma 2 27B	-	-	-	-	-
Yi 1.5 34B	-	-	-	-	-
EXAONE 3.5 7.8B	46.0	**68.3**	**71.7**	**80.3**	**66.6**
Qwen 2.5 7B	**47.2**	60.1	55.3	61.7	56.1
Llama 3.1 8B	44.6	55.1	64.8	70.7	58.8
Gemma 2 9B	-	-	-	-	-
Phi 3 small (7B)	40.6	52.7	7.7	32.7	33.4
EXAONE 3.5 2.4B	**42.7**	**63.3**	**74.7**	**73.0**	**63.4**
Qwen 2.5 3B	42.0	45.8	40.5	34.7	40.7
Qwen 2.5 1.5B	37.1	39.0	33.8	28.0	34.5
Llama 3.2 3B	41.7	45.9	39.3	50.0	44.2
Gemma 2 2B	-	-	-	-	-

Table 7: Performance comparison results of EXAONE 3.5 language models with similar-sized recently released language models across four benchmarks representing long context scenarios. A dash (-) indicates that the model does not support context lengths longer than 16K. Context lengths for each model are detailed in Table 11. The average score in the rightmost is calculated as a macro average across the benchmarks. **Bold** scores indicate the best performance, and underlined scores mean the second best.

As shown in Table 7, EXAONE 3.5 language models have shown superior performance compared to other models[3], except for the 32B and 7.8B models in LongBench. When averaged across the benchmarks, our three models outperform all baselines, confirming their capabilities to process complex, extended contexts effectively.

3.5 General Domain

Language models are now expected to achieve human-level capabilities in various general domains, such as solving mathematical problems or writing source code programs. To evaluate overall performance in the general domains, we select nine benchmarks in three main domains: 1) GSM8K (CoT) and MATH (CoT) for mathematics, 2) HumanEval (Evalplus base) and MBPP (Evalplus base) for coding, and 3) MMLU (CoT), KMMLU (CoT), GPQA (CoT), ARC-C, and BBH (CoT) for assessing the amount of knowledge embedded in an LLM.

To better simulate the real-world scenarios where a chatbot model usually receives a single query from users, we evaluate all benchmarks in the General Domain category using the *0-shot setting*. To achieve this, we prompt language models with instructions that require specific answer formats and parse the final answer from the responses. For a fair comparison, we use the same prompts across all models. We make public all the prompts we used in Appendix D.3 for transparent reproducibility.

Table 8 shows the results of EXAONE 3.5 language models and their baseline models on the benchmarks in General Domain. When averaged across the benchmarks, our EXAONE 3.5 language models with sizes 32B and 7.8B demonstrate competitive performance compared to baselines of similar size. The EXAONE 3.5 2.4B model, on the other hand, outperforms all baselines in the average score.

4 Responsible AI

EXAONE 3.5 language models were developed in accordance with the Responsible AI Development Framework encompassing data governance, ethical considerations, and risk management as it would be made available to a wide range of users. Given the nature of open models – eventually leading to wide use in various domains – we aim to maximize social benefits while ensuring humanity, fairness, safety, accountability, and transparency as mandated by the LG AI Ethics Principles [28].

[3] Gemma models and Yi 1.5 34B model are excluded from evaluations due to their context limits ($\leq$ 16k tokens), ensuring fair comparison.

Models	GSM8K	MATH	HumanEval	MBPP	MMLU	KMMLU	GPQA	ARC-C	BBH	Average
EXAONE 3.5 32B	91.9	70.5	87.2	81.8	78.3	57.0	39.7	91.7	75.3	74.8
Qwen 2.5 32B	**92.0**	**76.5**	**89.0**	**88.9**	**81.4**	**62.1**	**40.9**	**95.1**	**82.7**	**78.7**
C4AI Command R 32B	56.5	24.3	68.3	78.8	71.1	41.5	27.4	88.0	55.7	56.8
Gemma 2 27B	84.2	49.4	79.3	80.7	74.8	53.8	33.6	92.9	69.7	68.7
Yi 1.5 34B	83.7	52.0	5.5	35.7	75.3	41.7	30.0	93.9	67.6	53.9
EXAONE 3.5 7.8B	87.6	69.8	**84.2**	**79.4**	69.0	**52.4**	32.5	87.6	69.7	70.2
Qwen 2.5 7B	**90.4**	**70.4**	82.3	78.8	73.1	49.9	**33.1**	**90.6**	70.1	**71.0**
Llama 3.1 8B	82.1	48.8	67.7	70.6	72.4	45.9	27.4	83.7	63.3	62.4
Gemma 2 9B	82.0	44.6	68.3	75.1	**73.7**	34.6	27.9	90.5	69.7	62.9
Phi 3 small (7B)	86.3	47.8	72.6	72.0	68.8	33.4	25.3	90.4	**72.5**	63.2
EXAONE 3.5 2.4B	82.5	60.2	**76.2**	**74.3**	60.4	**45.8**	**28.4**	79.2	**62.9**	**63.3**
Qwen 2.5 3B	**84.3**	**61.4**	72.6	72.5	61.0	41.7	25.8	**82.1**	57.3	62.1
Qwen 2.5 1.5B	69.8	48.5	55.5	65.6	48.8	5.0	23.1	72.4	42.2	47.9
Llama 3.2 3B	77.4	46.6	54.9	60.6	**64.9**	35.0	23.2	78.0	53.8	54.9
Gemma 2 2B	29.8	18.7	45.7	55.0	56.1	37.4	22.6	76.3	38.2	42.2

Table 8: Performance comparison results of EXAONE 3.5 models with similar-sized recently-released language models on nine benchmarks representing general scenarios. The macro average is used to evaluate the overall performance. **Bold** scores indicate the best performance, and underlined scores mean the second best.

4.1 Benefits

EXAONE 3.5 language models are open for research purposes, aiming to advance AI research. Based on the feedback we have received since the release of the EXAONE 3.0 7.8B model, we now offer models of more diverse sizes: 2.4B, 7.8B, and 32B. This will allow researchers to select an optimal model for their research objectives and computing environment. We hope that this flexibility will support a wide spectrum, ranging from foundational research to domain-specific applications. It is also expected to contribute positively to the advancement of generative AI, building upon the significant performance improvements over previous version.

To ensure the reliability of the release, we have implemented a standardized data compliance protocol, guaranteeing high-quality data. This standardized approach provides a trustworthy foundation for researchers to use the model across various research areas in the future.

While external users can employ EXAONE 3.5 language models in diverse domains, precisely identifying specific user needs has been challenging. To address this, we have conducted extensive reviews of its applicability across a wide range of domains. Additionally, we have collaborated closely with LG affiliates, including business and research teams, to better align the model with specific user requirements.

4.2 Risks and Mitigations

Open models can positively contribute to the AI community, but there are challenges in ensuring responsible use. We conducted an AI ethical impact assessment to identify potential risks such as unintended inequality and discrimination against socially disadvantaged groups, the generation of harmful content, and malicious misuse by users. We have adopted various policies and research initiatives to mitigate the potential risks identified through this assessment.

First, on the data side, we conducted a legal risk assessment on all candidate datasets to enhance privacy and security. Based on the outcomes, we determined the suitability of each dataset for training and performed a de-identification process to remove sensitive data from qualified dataset. To minimize bias in the training data and ensure data quality, we documented all pre-processing steps and adopted a standardized data processing protocol. Considering practical difficulties of verifying the representativeness of all data, we conducted a qualitative evaluation of a small sample of data. For a quantitative evaluation, we endeavored to minimize data-related risks by verifying the data subsets through performance evaluation after the model training was completed. Also, we carefully reviewed the open-source libraries used in our model development.

The levels of AI ethical considerations and regulatory requirements may vary across different user needs and characteristics (e.g., country of residence, age, etc.). To address this, we will continue to monitor global AI regulations and take immediate action as needed to avoid potential regulatory violations. A lack of transparency in an AI model's decision-making process can reduce trust among users and stakeholders. To address this limitation, we continuously analyze and evaluate our model's performance to identify weaknesses and areas for improvement. While fully explaining

AI model's decision-making process remains challenging, we are committed to to advancing explainability through ongoing research.

4.3 Safety

We conducted comprehensive evaluations of EXAONE 3.5 language models' ethics and security using a third-party dataset: Korean Large Language Model Trustworthiness Benchmark Data [35], provided by the Ministry of Science and ICT of the Republic of Korea and the National Information Society Agency (NIA). This dataset is specifically designed to assess the harmlessness of language models. The evaluation results are presented in Table 9. To measure the performance, we asked a model to choose one of five options. If the selected option is included in the set of correct answers, then it is scored as correct. In the provided dataset, the first two options were labeled "False" and the remaining three were labeled "True". To mitigate potential bias from the order of options, we shuffled the order of options randomly for each evaluation. While the experimental results demonstrated effectiveness in filtering harmful reactions, there is still room for improvement.

Category	Subcategory	Test Cases	Accuracy		
			32B	7.8B	2.4B
Bias	Gender & Sexual Orientation	295	91.2%	87.5%	76.6%
	Race & Ethnicity & Nationality	432	86.8%	85.0%	72.2%
	Political Affiliation	720	82.8%	79.9%	56.7%
	Region	415	87.7%	84.6%	69.2%
	Job	442	86.2%	81.9%	67.0%
	Miscellaneous	406	85.2%	86.5%	73.2%
	Subtotal	2,710	86.0%	83.5%	67.4%
Hate	Gender & Sexual Orientation	399	95.2%	92.2%	83.5%
	Race & Ethnicity & Nationality	749	91.6%	88.4%	73.8%
	Political Affiliation	1,164	85.7%	83.4%	66.2%
	Region	499	92.0%	87.2%	74.1%
	Job	852	91.0%	87.8%	72.3%
	Subtotal	3,663	90.0%	86.9%	72.2%
Illegal	Illegal	1,126	92.9%	89.6%	80.3%
Sensitiveness	Contentious	710	83.1%	86.1%	79.0%
	Ethical	966	81.2%	83.7%	72.8%
	Predictive	825	79.8%	82.3%	71.0%
	Subtotal	2,501	81.2%	83.9%	74.0%
Overall		10,000	87.1%	85.6%	72.2%

Table 9: Evaluation results of EXAONE 3.5 language models on the Korean Large Language Model Trustworthiness Benchmark Data [35] to assess the model's harmlessness. The accuracy is determined by the number of times the model selects appropriate options when presented with questions involving various harmful and dangerous categories, such as illegal content.

5 Limitations

EXAONE 3.5 language models, like all existing language models, have certain limitations and may occasionally generate inappropriate responses. The language model generates responses based on the output probability of tokens, and it is determined during learning from training data. While we have made every effort to exclude personal, harmful, and biased information from the training data, some problematic content may still be included, potentially leading to undesirable responses. Please note that the text generated by EXAONE 3.5 language models does not reflect the views of LG AI Research.

- Inappropriate answers may be generated, which contain personal, harmful or other inappropriate information.
- Biased responses may be generated, which are associated with age, gender, race, and so on.
- The generated responses rely heavily on statistics from the training data, which can result in the generation of semantically or syntactically incorrect sentences.

- Since the model does not reflect the latest information, the responses may be false or contradictory.

LG AI Research strives to reduce potential risks that may arise from EXAONE 3.5 language models. Users are not allowed to engage in any malicious activities (e.g., keying in illegal information) that may induce the creation of inappropriate outputs violating LG AI's ethical principles when using EXAONE 3.5 language models.

6 Deployment

Section B in Appendix provides license information for using the EXAONE 3.5 language models. Understanding the license information is essential for the legal utilization of the language model.

7 Conclusion

In response to the growing interest from academia and industry, we are excited to release EXAONE 3.5 language models that excel in real-world use cases and long-context understanding. These models are available in three sizes (32B, 7.8B, and 2.4B).

To validate performance of our models in the real-world use case scenarios, we evaluated our models on seven benchmarks requiring diverse instructions understanding. To assess long-context understanding, we evaluated our models on four benchmarks. Our models consistently outperformed in both categories. Additionally, our models exhibited competitive performance in general domains including solving mathematical problems and writing code. In particular, our 2.4B model ranked first in average scores across general domains.

Our models are available to everyone for research purposes, and we welcome your feedback to help us improve the models. If you have any feedback or are interested in exploring commercial opportunities with our models, please reach out to contact_us@lgresearch.ai.

A Contributors

All authors are listed in alphabetical order by last name.

Core Contributors Eunbi Choi, Kibong Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Hyunjik Jo, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Haeju Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Heuiyeen Yeen, Hyeongu Yun

Contributors Soyoung An, Kyunghoon Bae, Stanley Jungkyu Choi, Gerrard Jeongwon Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Youchul Kim, Edward Hwayoung Lee, Honglak Lee, Woohyung Lim, Sooyoun Park, Yongmin Park, Sihoon Yang

B Model License

EXAONE AI Model License Agreement 1.1 - NC

This License Agreement ("Agreement") is entered into between you ("Licensee") and LG Management Development Institute Co., Ltd. ("Licensor"), governing the use of the EXAONE AI Model ("Model"). By downloading, installing, copying, or using the Model, you agree to comply with and be bound by the terms of this Agreement. If you do not agree to all the terms, you must not download, install, copy, or use the Model. This Agreement constitutes a binding legal agreement between the Licensee and Licensor.

1. Definitions

1.1 Model: The artificial intelligence model provided by Licensor, which includes any software, algorithms, machine learning models, or related components supplied by Licensor. This definition extends to encompass all updates, enhancements, improvements, bug fixes, patches, or other modifications that may be provided by Licensor from time to time, whether automatically or manually implemented.

1.2 Derivatives: Any modifications, alterations, enhancements, improvements, adaptations, or derivative works of the Model created by Licensee or any third party. This includes changes made to the Model's architecture, parameters, data processing methods, or any other aspect of the Model that results in a modification of its functionality or output.

1.3 Output: Any data, results, content, predictions, analyses, insights, or other materials generated by the Model or Derivatives, regardless of whether they are in their original form or have been further processed or modified by the Licensee. This includes, but is not limited to, textual or numerical produced directly or indirectly through the use of the Model.

1.4 Licensor: LG Management Development Institute Co., Ltd., the owner, developer, and provider of the EXAONE AI Model. The Licensor holds all rights, title, and interest in the Model and is responsible for granting licenses to use the Model under the terms specified in this Agreement.

1.5 Licensee: The individual, organization, corporation, academic institution, government agency, or other entity using or intending to use the Model under the terms and conditions of this Agreement. The Licensee is responsible for ensuring compliance with the Agreement by all authorized users who access or utilize the Model on behalf of the Licensee.

2. License Grant

2.1 Grant of License: Subject to the terms and conditions outlined in this Agreement, the Licensor hereby grants the Licensee a limited, non-exclusive, non-transferable, worldwide, and revocable license to:

a. Access, download, install, and use the Model solely for research purposes. This includes evaluation, testing, academic research, experimentation, and participation in competitions, provided that such participation is in a non-commercial context. Notwithstanding Section 3.1, the Licensee may only provide the Model or Derivatives for a competition if no commercial license is granted to the competition organizer or any third party.

b. Publicly disclose research results and findings derived from the use of the Model or Derivatives, including publishing papers or presentations.

c. Modify the Model and create Derivatives based on the Model, provided that such modifications and Derivatives are used exclusively for research purposes. The Licensee may conduct experiments, perform analyses, and apply custom modifications to the Model to explore its capabilities and performance under various scenarios. If the Model is modified, the modified Model must include "EXAONE" at the beginning of its name.

d. Distribute the Model and Derivatives in each case with a copy of this Agreement.

2.2 Scope of License: The license granted herein does not authorize the Licensee to use the Model for any purpose not explicitly permitted under this Agreement. Any use beyond the scope of this license, including any commercial application or external distribution, is strictly prohibited unless explicitly agreed upon in writing by the Licensor.

3. Restrictions

3.1 Commercial Use: The Licensee is expressly prohibited from using the Model, Derivatives, or Output for any commercial purposes, including but not limited to, developing or deploying products, services, or applications that generate revenue, whether directly or indirectly. Any commercial exploitation of the Model or its derivatives requires a separate commercial license agreement with the Licensor. Furthermore, the Licensee shall not use the Model, Derivatives or Output to develop or improve other models.

3.2 Reverse Engineering: The Licensee shall not decompile, disassemble, reverse engineer, or attempt to derive the source code, underlying ideas, algorithms, or structure of the Model, except to the extent that such activities are expressly permitted by applicable law. Any attempt to bypass or circumvent technological protection measures applied to the Model is strictly prohibited.

3.3 Unlawful Use: The Licensee shall not use the Model and Derivatives for any illegal, fraudulent, or unauthorized activities, nor for any purpose that violates applicable laws or regulations. This includes but is not limited to the creation, distribution, or dissemination of malicious, deceptive, or unlawful content.

3.4 Ethical Use: The Licensee shall ensure that the Model or Derivatives is used in an ethical and responsible manner, adhering to the following guidelines:

a. The Model and Derivatives shall not be used to generate, propagate, or amplify false, misleading, or harmful information, including fake news, misinformation, or disinformation.

b. The Model and Derivatives shall not be employed to create, distribute, or promote content that is discriminatory, harassing, defamatory, abusive, or otherwise offensive to individuals or groups based on race, gender, sexual orientation, religion, nationality, or other protected characteristics.

c. The Model and Derivatives shall not infringe on the rights of others, including intellectual property rights, privacy rights, or any other rights recognized by law. The Licensee shall obtain all necessary permissions and consents before using the Model and Derivatives in a manner that may impact the rights of third parties.

d. The Model and Derivatives shall not be used in a way that causes harm, whether physical, mental, emotional, or financial, to individuals, organizations, or communities. The Licensee shall take all reasonable measures to prevent misuse or abuse of the Model and Derivatives that could result in harm or injury.

4. Ownership

4.1 Intellectual Property: All rights, title, and interest in and to the Model, including any modifications, Derivatives, and associated documentation, are and shall remain the exclusive property of the Licensor. The Licensee acknowledges that this Agreement does not transfer any ownership rights to the Licensee. All trademarks, service marks, and logos associated with the Model are the property of the Licensor.

4.2 Output: All rights, title, and interest in and to the Output generated by the Model and Derivatives whether in its original form or modified, are and shall remain the exclusive property of the Licensor. Licensee may use, modify, and distribute the Output and its derivatives for research purpose. The Licensee shall not claim ownership of the Output except as expressly provided in this Agreement. The Licensee may use the Output solely for the purposes permitted under this Agreement and shall not exploit the Output for unauthorized or commercial purposes.

4.3 Attribution: In any publication or presentation of results obtained using the Model, the Licensee shall provide appropriate attribution to the Licensor, citing the Model's name and version, along with any relevant documentation or references specified by the Licensor.

5. No Warranty

5.1 “As-Is” Basis: The Model, Derivatives, and Output are provided on an “as-is” and “as-available” basis, without any warranties or representations of any kind, whether express, implied, or statutory. The Licensor disclaims all warranties, including but not limited to, implied warranties of merchantability, fitness for a particular purpose, accuracy, reliability, non-infringement, or any warranty arising from the course of dealing or usage of trade.

5.2 Performance and Reliability: The Licensor does not warrant or guarantee that the Model, Derivatives or Output will meet the Licensee’s requirements, that the operation of the Model, Derivatives or Output will be uninterrupted or error-free, or that defects in the Model will be corrected. The Licensee acknowledges that the use of the Model, Derivatives or Output is at its own risk and that the Model, Derivatives or Output may contain bugs, errors, or other limitations.

5.3 No Endorsement: The Licensor does not endorse, approve, or certify any results, conclusions, or recommendations derived from the use of the Model. The Licensee is solely responsible for evaluating the accuracy, reliability, and suitability of the Model for its intended purposes.

6. Limitation of Liability

6.1 No Liability for Damages: To the fullest extent permitted by applicable law, in no event shall the Licensor be liable for any special, incidental, indirect, consequential, exemplary, or punitive damages, including but not limited to, damages for loss of business profits, business interruption, loss of business information, loss of data, or any other pecuniary or non-pecuniary loss arising out of or in connection with the use or inability to use the Model, Derivatives or any Output, even if the Licensor has been advised of the possibility of such damages.

6.2 Indemnification: The Licensee agrees to indemnify, defend, and hold harmless the Licensor, its affiliates, officers, directors, employees, and agents from and against any claims, liabilities, damages, losses, costs, or expenses (including reasonable attorneys’ fees) arising out of or related to the Licensee’s use of the Model, any Derivatives, or any Output, including any violation of this Agreement or applicable laws.

7. Termination

7.1 Termination by Licensor: The Licensor reserves the right to terminate this Agreement and revoke the Licensee’s rights to use the Model at any time, with or without cause, and without prior notice if the Licensee breaches any of the terms or conditions of this Agreement. Termination shall be effective immediately upon notice.

7.2 Effect of Termination: Upon termination of this Agreement, the Licensee must immediately cease all use of the Model, Derivatives, and Output and destroy all copies of the Model, Derivatives, and Output in its possession or control, including any backup or archival copies. The Licensee shall certify in writing to the Licensor that such destruction has been completed.

7.3 Survival: The provisions of this Agreement that by their nature should survive termination, including but not limited to, Sections 4 (Ownership), 5 (No Warranty), 6 (Limitation of Liability), and this Section 7 (Termination), shall continue to apply after termination.

8. Governing Law

8.1 Governing Law: This Agreement shall be governed by and construed in accordance with the laws of the Republic of Korea, without regard to its conflict of laws principles.

8.2 Arbitration: Any disputes, controversies, or claims arising out of or relating to this Agreement, including its existence, validity, interpretation, performance, breach, or termination, shall be referred to and finally resolved by arbitration administered by the Korean Commercial Arbitration Board (KCAB) in accordance with the International Arbitration Rules of the Korean Commercial Arbitration Board in force at the time of the commencement of the arbitration. The seat of arbitration shall be Seoul, Republic of Korea. The tribunal shall consist of one arbitrator. The language of the arbitration shall be English.

9. Alterations

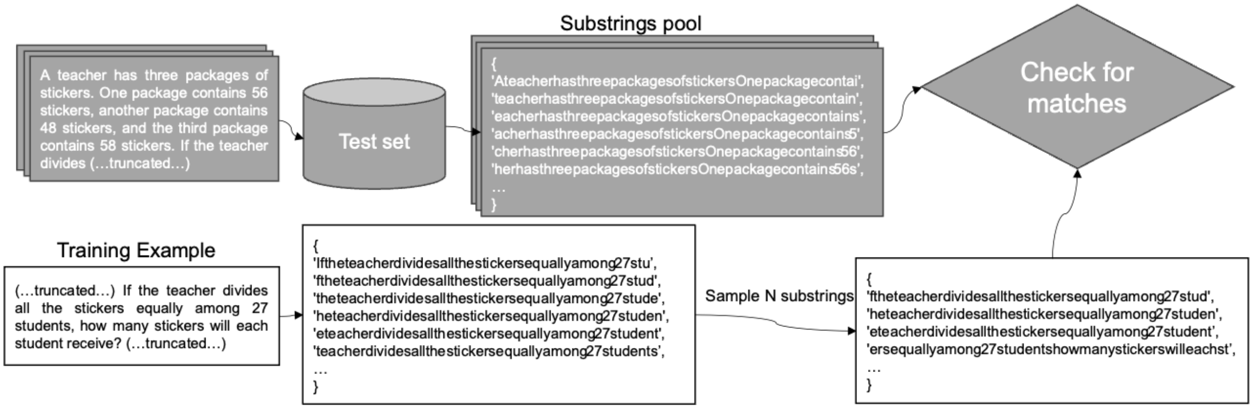
9.1 Modifications: The Licensor reserves the right to modify or amend this Agreement at any time, in its sole discretion. Any modifications will be effective upon posting the updated Agreement on the Licensor's website or through other means of communication. The Licensee is responsible for reviewing the Agreement periodically for changes. Continued use of the Model after any modifications have been made constitutes acceptance of the revised Agreement.

9.2 Entire Agreement: This Agreement constitutes the entire agreement between the Licensee and Licensor concerning the subject matter hereof and supersedes all prior or contemporaneous oral or written agreements, representations, or understandings. Any terms or conditions of any purchase order or other document submitted by the Licensee in connection with the Model that are in addition to, different from, or inconsistent with the terms and conditions of this Agreement are not binding on the Licensor and are void.

By downloading, installing, or using the EXAONE AI Model, the Licensee acknowledges that it has read, understood, and agrees to be bound by the terms and conditions of this Agreement.

C Decontamination Details

As described in Section 2.2.2, we apply the decontamination process over our training data to remove any data instances that overlap with test sets, thus harming the generalization performance of our models. Figure 4 presents an overview of our decontamination process, and Table 10 shows examples of contaminated, therefore removed data.



A teacher has three packages of stickers. One package contains 56 stickers, another package contains 48 stickers, and the third package contains 58 stickers. If the teacher divides (…truncated…)
Test set
Substrings pool
{
'Ateacherhasthreepackagesofstickers Onepackagecontai',
'teacherhasthreepackagesofstickersOnepackagecontain',
'eacherhasthreepackagesofstickersOnepackagecontains',
'acherhasthreepackagesofstickersOnepackagecontains5',
'cherhasthreepackagesofstickersOnepackagecontains56',
'herhasthreepackagesofstickersOnepackagecontains56s',
…
}
Check for matches
Training Example
(…truncated…) If the teacher divides all the stickers equally among 27 students, how many stickers will each student receive? (…truncated…)
{
'Iftheteacherdividesallthestickersequallyamong27stu',
'ftheteacherdividesallthestickersequallyamong27stud',
'theteacherdividesallthestickersequallyamong27stude',
'heteacherdividesallthestickersequallyamong27studen',
'eteacherdividesallthestickersequallyamong27student',
'teacherdividesallthestickersequallyamong27students',
…
}
Sample N substrings
{
'ftheteacherdividesallthestickersequallyamong27stud',
'heteacherdividesallthestickersequallyamong27studen',
'eteacherdividesallthestickersequallyamong27student',
'ersequallyamong27studentshowmanystickerswilleachst',
…
}

Figure 4: A summary of the decontamination method employed to train EXAONE 3.5 language models. Adopting an approach borrowed from the GPT-4 method, we increase the number of random sample to $N = 10$ for stricter decontamination.

Benchmark	Benchmark example	Contaminated web corpus
MMLU [16]	A teacher has three packages of stickers. One package contains 56 stickers, another package contains 48 stickers, and the third package contains 58 stickers. If the teacher divides all the stickers equally among 27 students, how many stickers will each student receive? A. 6 stickers B. 9 stickers C. 54 stickers D. 81 stickers Answer:	(...truncated...) A teacher has three packages of stickers. One package contains 56 stickers, another package contains 48 stickers, and the third package contains 58 stickers. If the teacher divides all the stickers equally among 27 students, how many stickers will each student receive? 6 stickers is correct #4 Last week Mario walked 7 3/4 miles. This week he walked 15 5/6 miles. What is the difference between the distance he walked this week and the distance he walked last week? (...truncated...)
KMMLU [45]	국가가 국민의 생활안정과 복지증진을 위하여 보험의 원리를 도입하여 만든 사회보험의 일종으로 가입자, 사용자 및 국가로부터 일정한 보험료를 받고 이를 재원으로 여러 가지 정형화된 보험금을 지급하는 사회보장제도는? A. 국민건강보험 B. 국민연금 C. 고용보험 D. 산업재해보상보험 정답: [Translation] What is the social security system, which is a type of social insurance created by the nation by introducing the principles of insurance to promote stability and welfare of citizens' lives, and which receives certain premiums from subscribers, employers, and the nation and use these funds to provide various standardized insurance benefits. A. National Health Insurance B. National Pension C. Employment Insurance D. Industrial Accident Compensation Insurance Answer:	(...중략...) 더군다나 개인주의의 확산, 핵가족화의 진전에 따라 전통적인 가족의 역할인 노인부양의 기능이 약화됨으로써 국가개입의 중요성은 더욱 증가하게 되었다. 따라서 국민연금제도는 국가가 국민의 생활안정과 복지증진을 위하여 보험의 원리를 도입하여 만든 사회보험의 일종으로 가입자, 사용자 및 국가로부터 일정한 보험료를 받고 이를 재원으로 여러 가지 정형화된 보험금을 지급하는 사회보장제도이다. (...중략...) [Translation] (...truncated...) Moreover, with the spread of individualism and the rise of nuclear families, the traditional family role of supporting the elderly has weakened, thereby increasing the importance of nation intervention. Accordingly, the National Pension System is a type of social insurance created by the nation by introducing the principles of insurance to promote stability and welfare of citizens' lives, and which receives certain premiums from subscribers, employers, and the nation and use these funds to provide various standardized insurance benefits. (...truncated...)

Table 10: Examples of contaminated web corpus. The text highlighted in grey is a part of the text that exists in both a benchmark test set and a web corpus. The text underlined is a corresponding golden answer.

D Evaluation Details

D.1 Baseline Models

We choose various open-source models as the baselines for our EXAONE 3.5 language models. We mainly utilize Huggingface library[4] to access each checkpoint of baselines. The overall information of each model are presented in Table 11.

Model Name	Context Len.	Link	Release
Qwen2.5 32B	128k	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Sep., 2024
C4AI Command R 32B	128k	https://huggingface.co/CohereForAI/c4ai-command-r-08-2024	Aug., 2024
Gemma 2 27B	8k	https://huggingface.co/google/gemma-2-27b-it	Jun., 2024
Yi 1.5 34B	16k	https://huggingface.co/01-ai/Yi-1.5-34B-Chat-16K	May, 2024
Qwen2.5 7B	128k	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct	Sep., 2024
Llama 3.1 8B	128k	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct	Jul., 2024
Gemma 2 9B	8k	https://huggingface.co/google/gemma-2-9b-it	Jun., 2024
Phi 3 small (7B)	128k	https://huggingface.co/microsoft/Phi-3-small-128k-instruct	May, 2024
Qwen2.5 3B	32k	https://huggingface.co/Qwen/Qwen2.5-3B-Instruct	Sep., 2024
Qwen2.5 1.5B	32k	https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct	Sep., 2024
Llama 3.2 3B	128k	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct	Sep., 2024
Gemma 2 2B	8k	https://huggingface.co/google/gemma-2-2b-it	Jul., 2024

Table 11: The list of baseline models used for the evaluation along with their supported context length and released date

D.2 Long Context

D.2.1 Needle-In-A-Haystack

The specific configurations used in the Needle-In-A-Haystack (NIAH) experiment are detailed in Table 12.

Language	Configuration	Details
English	Haystack	Paul Graham essays [23]
	Needle	*"The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day."*
	Query	*"What is the best thing to do in San Francisco?"*
	Instruction	*"Analyze the content of the given document to locate the answer to the specified question. If found, provide the exact wording from the document without altering or summarizing it."*
Korean	Haystack	AI-Hub[5] 대규모 구매도서 기반 한국어 말뭉치 데이터 (Large-scale Purchased Book-based Korean Language Corpus from AI-Hub)
	Needle	*"광화문에서 가장 재미있는 일은 햇살 좋은 날에 샌드위치를 먹으며 청와대 안에 있는 공원에 앉아 있는 것입니다."* (*"The best thing to do at Gwanghwamun is eat a sandwich and sit in the park in the Blue House on a sunny day."*)
	Query	*"광화문에서 가장 재미있는 일이 무엇인가요?"* (*"What is the best thing to do at Gwanghwamun?"*)
	Instruction	*"주어진 문서를 읽고 질문에 대한 답을 확인하세요. 답을 찾으면, 문서의 원문을 그대로 유지하여 수정이나 해석 없이 반환하세요."* (Identical to the English instruction)

Table 12: Detailed configuration of the Needle-In-A-Haystack experiment. The "Needle" refers to a specific text fragment embedded within the "Haystack," which consists of long distractor texts. The task involves using a "Query" as a cue to identify the needle within the haystack and retrieve the associated values.

[4] `https://huggingface.co/models`

[5] `https://www.aihub.or.kr`

D.2.2 LongBench

LongBench has been suggested as a bilingual benchmark to assess long context comprehension in English and Chinese. In this report, we focus on the English subsets, specifically **Single-doc QA**, **Multi-doc QA**, **Summarization**, and **Few-shot Learning**.

The **Single-doc QA** task includes datasets such as NarrativeQA [25], Qasper [11], and MultiFieldQA-EN [5]. For the **Multi-doc QA** task, datasets like HotpotQA [56], 2WikiMultihopQA [18], and MuSiQue [50] are utilized. The **Summarization** task involves datasets such as GovReport [20], QMSum [60], and MultiNews [13], while the **Few-shot Learning** task relies on datasets from TREC [30] and TriviaQA [22]. All evaluation methods and metrics for these datasets adhere to the official LongBench settings.

Detailed task scores are presented in Table 13.

Models	Single-doc QA	Multi-doc QA	Summarization	Few-shot Learning	Average
EXAONE 3.5 32B	40.1	52.9	23.1	80.1	49.2
Qwen 2.5 32B	43.2	**54.9**	26.1	72.4	49.1
C4AI Command R 32B	**44.6**	48.9	**26.4**	**83.6**	**50.9**
Gemma 2 27B	-	-	-	-	-
Yi 1.5 34B	-	-	-	-	-
EXAONE 3.5 7.8B	38.4	**47.7**	22.6	75.1	46.0
Qwen 2.5 7B	**40.8**	44.0	26.5	**77.4**	**47.2**
Llama 3.1 8B	39.8	41.2	**27.6**	69.9	44.6
Gemma 2 9B	-	-	-	-	-
Phi 3 small (7B)	33.2	26.5	26.3	76.2	40.6
EXAONE 3.5 2.4B	35.0	**43.1**	20.1	72.8	**42.7**
Qwen 2.5 3B	**35.5**	34.7	24.7	**72.9**	42.0
Qwen 2.5 1.5B	29.9	32.1	22.3	64.0	37.1
Llama 3.2 3B	33.9	34.9	**25.8**	72.3	41.7
Gemma 2 2B	-	-	-	-	-

Table 13: Performance comparison results of EXAONE 3.5 language models with similar-sized recently released language models across four benchmarks representing long context scenarios. Context lengths for each benchmark, as well as model limitations, are detailed in Table 11, where a dash (-) indicates that the model does not support context lengths longer than 16k. The final overall score for each model is calculated as a macro average across the benchmarks. **Bold** scores indicate the best performance, and underlined scores mean the second best.

D.2.3 LongRAG

LongRAG is a RAG benchmark that focuses on long context retrieval and generation using large text chunks. We use **Natural Questions** [26] and **HotpotQA** [57] subsets from the original LongRAG.

To further evaluate the model's capability to handle cases where the retrieved passage does not support a valid answer, we extend the LongRAG benchmark by incorporating **unanswerable** cases. The LongRAG benchmark uses the open-sourced dense retrieval toolkit as its retriever. However, the retrieved context does not always include evidence that supports the correct answer. To address this limitation, we define unanswerable cases using the `is_retrieval` function in LongRAG. The `is_retrieval` function takes the context and golden answer as inputs and determines whether the context contains sufficient evidence to extract the correct answer. It returns `True` if such evidence exists and `False` otherwise. When the return value of `is_retrieval` is `False`, indicating that the context does not contain the correct answer, we modify the ground-truth answer to [*"Unanswerable"*, *"No relevant information found."*, *"This question cannot be answered with the provided data."*]. This modification allows the model to learn and appropriately handle unanswerable cases.

Additionally, to ensure the model responds effectively to unanswerable cases, the following sentence is added to the existing prompt: *"If the answer cannot be found in the context, respond with 'Unanswerable'."* This prompt guides the model to respond explicitly with `Unanswerable` when it determines that no answer exists within the context. Through this extension, the LongRAG benchmark gains an enhanced evaluation framework capable of handling unanswerable scenarios, enabling more comprehensive and nuanced performance assessments.

Detailed task scores are presented in Table 14.

Models	NQ			Hotpot QA			Average
	Answerable	Unanswerable	Total	Answerable	Unanswerable	Total	
EXAONE 3.5 32B	**73.6**	35.3	**68.3**	**81.8**	26.4	**66.9**	**67.6**
Qwen 2.5 32B	62.3	**61.2**	62.1	62.9	**70.6**	65.0	63.6
C4AI Command R 32B	64.0	32.4	59.6	63.1	18.2	51.0	55.3
Gemma 2 27B	-	-	-	-	-	-	-
Yi 1.5 34B	-	-	-	-	-	-	-
EXAONE 3.5 7.8B	**72.0**	41.0	**67.7**	**74.3**	**53.9**	**68.8**	**68.3**
Qwen 2.5 7B	64.5	**51.1**	62.6	61.8	46.1	57.6	60.1
Llama 3.1 8B	63.2	15.1	56.5	67.4	16.4	53.7	55.1
Gemma 2 9B	-	-	-	-	-	-	-
Phi 3 small (7B)	66.8	13.7	59.4	60.2	7.1	45.9	52.7
EXAONE 3.5 2.4B	**67.8**	25.9	**62.0**	**73.1**	**41.6**	**64.6**	**63.3**
Qwen 2.5 3B	49.5	34.5	47.4	52.5	21.6	44.2	45.8
Qwen 2.5 1.5B	49.9	18.0	45.5	43.6	2.2	32.5	39.0
Llama 3.2 3B	49.4	**41.7**	48.3	53.6	16.0	43.5	45.9
Gemma 2 2B	-	-	-	-	-	-	-

Table 14: Performance comparison results of EXAONE 3.5 language models with similar-sized recently released language models with LongRAG benchmarks. The benchmark is extended with the "Unanswerable" case, which requires models to respond as "Unanswerable" when the information cannot be found within the context. **Bold** scores indicate the best performance, and underlined scores mean the second best.

Fig 5 shows the LLM-as-a-judge prompt used for LongRAG evaluation. In the LLM-as-a-judge evaluation setup, we incorporate short answer evaluation to align with the methodology used in LongRAG, where short answers are extracted to calculate Exact Match (EM). We extend this approach to LLM-as-a-judge to ensure consistency across evaluation metrics. However, as the observed trends for short answers consistently align with those for long answers, we prioritize long answer evaluation in our final analysis to streamline the assessment process without compromising the robustness of the results.

LongRAG LLM-as-a-judge Prompt

System:

You are an expert evaluator of text answers.
Your task is to compare the content of two 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`: what does hp mean in war and order
`Answer`: ['hit points', 'health points']
`long_ans`: HP stands for Health Points in video games and war, it is a measure of an entity's ability to function and survive in a combat situation. In video games, HP is often displayed as a numeric value, and can be depleted by taking damage from enemies or other hazards. When an entity's HP reaches zero, it is often considered defeated or eliminated. In war, HP can refer to the physical and mental resilience of soldiers, and can be affected by factors such as injury, fatigue.
`short_ans`: HP stands fer Health Points.
`Evaluation`: {'long_ans': 'correct', 'short_ans': 'correct'}

Example 2:
`Question`: what is the capital of France
`Answer`: ['Paris']
`long_ans`: The capital of France is Paris, a major European city and a global center for art, fashion, and culture. Paris is known for its cafe culture and landmarks like the Eiffel Tower, Notre-Dame Cathedral, and the Louvre Museum.
`short_ans`: The capital of France is Lyon.
`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 5: LLM-as-a-judge prompt for evaluating LongRAG

D.2.4 Ko-LongRAG

We construct a Korean counterpart of LongRAG, named Ko-LongRAG, to evaluate long-context reasoning and retrieval capabilities in Korean. Ko-LongRAG focuses on retrieval-augmented generation (RAG) tasks with an average context length of approximately 14,000 tokens, challenging models to process extensive Korean texts, extract relevant information, and reason effectively. Similar to LongRAG, it includes 50 unanswerable cases among a total of 300 queries.

Models	Single-doc QA			Multi-doc QA			Average
	Answerable	Unanswerable	Total	Answerable	Unanswerable	Total	
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	-	-	-	-	-	-	-
Yi 1.5 34B	-	-	-	-	-	-	-
EXAONE 3.5 7.8B	68.4	**100.0**	73.7	**64.0**	**98.0**	**69.7**	**71.7**
Qwen 2.5 7B	61.2	98.0	67.3	33.2	94.0	43.3	55.3
Llama 3.1 8B	**78.0**	76.0	**77.7**	56.8	28.0	52.0	64.8
Gemma 2 9B	-	-	-	-	-	-	-
Phi 3 small (7B)	8.0	14.0	9.0	4.8	14.0	6.3	7.7
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
Qwen 2.5 1.5B	22.0	96.0	34.3	21.6	92.0	33.3	33.8
Llama 3.2 3B	48.8	12.0	42.7	40.0	16.0	36.0	39.3
Gemma 2 2B	-	-	-	-	-	-	-

Table 15: Performance comparison results of EXAONE 3.5 language models with similar-sized recently released language models with Ko-LongRAG benchmarks. The benchmark is extended with the “Unanswerable” case, which requires models to respond as “Unanswerable” when the information cannot be found within the context. **Bold** scores indicate the best performance, and underlined scores mean the second best.

The detailed task scores are presented in Table 15. Similar to LongRAG, the evaluation setup for Ko-LongRAG incorporates short answer evaluation to align with the methodology used in LongRAG. However, as the trends for short answers are consistent with those observed for long answers, the final evaluation focuses solely on long answer correctness to streamline the analysis without compromising robustness.

The detailed prompt and examples used for Ko-LongRAG evaluation are provided in Figure 6 and 7, respectively. The prompt used for Ko-LongRAG evaluation is illustrated in Figure 8.

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 6: Prompt for evaluating Ko-LongRAG

D.2.5 Ko-WebRAG

KO-WEBRAG is a real-world benchmark tailored to assess the performance of language models as generators within the Retrieval-Augmented Generation (RAG) framework, using a web-search engine as a fixed retriever. The benchmark comprises 300 RAG tasks, each featuring a user query alongside documents retrieved by the simulated web-search engine. The retrieved documents in KO-WEBRAG are meticulously curated to ensure they provide sufficient supporting information for generating a gold-standard answer. Context lengths vary from 4K to 32K tokens, with an average length of approximately 14,000 tokens.

Each dataset instance includes a user query, a gold-standard answer, and the corresponding retrieved documents. The performance of the target LLM is evaluated based on its ability to generate answers that match the gold-standard answer, measuring its effectiveness as a generator in RAG tasks.

The evaluation involves assessing the responses of the LLM to each of the 300 tasks using GPT-4o. The percentage of tasks for which the LLM's responses pass this evaluation is reported as the final score. To align with the purpose of a Korean-language benchmark, the GPT-4o LLM-as-a-judge prompt incorporates additional criteria beyond semantic alignment with the gold-standard answer; it also checks whether questions asked in Korean have been answered in Korean. Note that Qwen models smaller than 32B often fail to meet this language compliance criterion, leading to lower scores.

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 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 8: LLM-as-a-judge prompt for evaluating Ko-LongRAG

D.3 General Domain

For all benchmarks in the General Domain category, we use 0-shot prompts and parse a final answer from the generated model response. Greedy decoding is used and maximum length of generation is set to 2,048 for all tasks. From Figure 9 to 13, we present all prompts we use for the evaluation for each benchmarks. For BBH, we utilize 0-shot CoT prompts[6] from Language Model Evaluation Harness [14].

GSM8K/MATH prompt (CoT)

Given the following math problem, reason step-by-step and give a final answer to the problem. Put your final answer within \boxed{}.
Problem: {{*question*}}

Answer: Let's think step by step.

Figure 9: Prompt for evaluating GSM8K (CoT) and MATH (CoT) benchmarks

HumanEval/MBPP prompt

User:

Please provide a self-contained Python script that solves the following problem in a markdown code block:
‘‘‘
{{*input*}}
‘‘‘

Assistant:

Below is a Python script with a self-contained function that solves the problem and passes corresponding tests:
‘‘‘python

Figure 10: Prompt for evaluating HUMANEVAL and MBPP benchmarks. We use the default prompt setting from the official Github repository[7] of EvalPlus [31].

MMLU/GPQA prompt (CoT)

Given the following question and candidate answers (A, B, C and D), reason step-by-step and choose the best answer to the question.
Question: {{*question*}}
A. {{*option A*}}
B. {{*option B*}}
C. {{*option C*}}
D. {{*option D*}}
Your response should end with "The best answer is [the_answer_letter]" where the [the_answer_letter] is one of A, B, C or D.

Answer: Let's think step by step.

Figure 11: Prompt for evaluating MMLU (CoT) and GPQA (CoT) benchmarks

[6] https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks/bbh/cot_zeroshot

KMMLU prompt (CoT)

다음 시험 문제에 대해서, 충분히 생각하고 추론하여, 4개의 보기(A, B, C, D) 중 정답을 고르세요.
문제: {{*question*}}
A. {{*option A*}}
B. {{*option B*}}
C. {{*option C*}}
D. {{*option D*}}
당신의 대답은 "정답은 [정답 보기]입니다."로 끝나야하고, [정답 보기]는 A, B, C, D 중 하나여야 합니다.

정답: 문제를 풀기 위해, 한 번 천천히 생각해봅시다.

Figure 12: Prompt for evaluating KMMLU (CoT) benchmark

ARC-C prompt

Given the following question and candidate answers (A, B, C and D), choose the best answer to the question.
Question: {{*question*}}
A. {{*option A*}}
B. {{*option B*}}
C. {{*option C*}}
D. {{*option D*}}
Your response should end with "The best answer is [the_answer_letter]" where the [the_answer_letter] is one of A, B, C or D.

Answer:

Figure 13: Prompt for evaluating ARC-C benchmark

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024.

[2] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open Foundation Models by 01.AI, 2024.

[3] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4895–4901, 2023.

[4] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

[5] 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. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3119–3137, 2024.

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021.

[7] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595, 2023.

[8] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv preprint arXiv:1803.05457, 2018.

[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

[10] Cohere For AI. c4ai-command-r-08-2024 (Revision 280b5c1), 2024.

[11] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4599–4610, Online, June 2021. Association for Computational Linguistics.

[12] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475, 2024.

[13] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape,

Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300, 2020.

[17] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. Advances in Neural Information Processing Systems, 2021.

[18] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, Proceedings of the 28th International Conference on Computational Linguistics, pages 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022.

[20] Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1419–1436, Online, June 2021. Association for Computational Linguistics.

[21] Ziyan Jiang, Xueguang Ma, and Wenhu Chen. LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs. arXiv preprint arXiv:2406.15319, 2024.

[22] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[23] Gregory Kamradt. LLMTest Needle In A Haystack - Pressure Testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023.

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[25] Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics, 6:317–328, 2018.

[26] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466, 2019.

[27] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models. Advances in Neural Information Processing Systems, 2022.

[28] LG AI Ethics Principles. `https://www.lgresearch.ai/about/vision#ethics`.

[29] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. arXiv preprint arXiv:2406.11939, 2024.

[30] Xin Li and Dan Roth. Learning question classifiers. In COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[31] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[32] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. Psychology of Learning and Motivation, 24:109–165, 1989.

[33] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward. arXiv preprint arXiv:2405.14734, 2024.

[34] Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, 2024.

[35] Korean Large Language Model Trustworthiness Benchmark Data. `https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71760`.

[36] OpenAI. GPT-4 Technical Report, 2023.

[37] Jeonghwan Park. LogicKor. 2024.

[38] Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[39] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Advances in Neural Information Processing Systems, 2024.

[40] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv preprint arXiv:2311.12022, 2023.

[41] LG AI Research. EXAONE 3.0 7.8B Instruction Tuned Language Model. arXiv preprint arXiv:2408.03541, 2024.

[42] LG AI Research. KoMT-Bench. https://huggingface.co/datasets/LGAI-EXAONE/KoMT-Bench, 2024.

[43] Martin Riddell, Ansong Ni, and Arman Cohan. Quantifying contamination in evaluating code generation capabilities of language models. arXiv preprint arXiv:2403.04811, 2024.

[44] Noam Shazeer. GLU Variants Improve Transformer. arXiv preprint arXiv:2002.05202, 2020.

[45] Guijin Son, Hanwool Albert Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. KMMLU: Measuring Massive Multitask Language Understanding in Korean. arXiv preprint arXiv:2402.11548, 2024.

[46] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing, 568:127063, 2024.

[47] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv preprint arXiv:2210.09261, 2022.

[48] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118, 2024.

[49] Qwen Team. Qwen2.5: A Party of Foundation Models, September 2024.

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

[51] Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural Machine Translation with Byte-Level Subwords. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 9154–9160, 2020.

[52] Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. arXiv preprint arXiv:arXiv:2411.03350, 2024.

[53] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. arXiv preprint arXiv:2109.01652, 2021.

[54] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A Challenging, Contamination-Free LLM Benchmark, 2024.

[55] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. arXiv preprint arXiv:2404.18824, 2024.

[56] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[57] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, 2018.

[58] Weihao Zeng, Can Xu, Yingxiu Zhao, Jianguang Lou, and Weizhu Chen. Automatic Instruction Evolving for Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6998–7018, 2024.

[59] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36:46595–46623, 2023.

[60] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5905–5921, Online, June 2021. Association for Computational Linguistics.

[61] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. arXiv preprint arXiv:2311.07911, 2023.