EXAONE 4.0: Unified Large Language Models Integrating Non-reasoning and Reasoning Modes

LG AI Research*

Abstract

This technical report introduces EXAONE 4.0, which integrates a NON-REASONING mode and a REASONING mode to achieve both the excellent usability of EXAONE 3.5 and the advanced reasoning abilities of EXAONE Deep. To pave the way for the agentic AI era, EXAONE 4.0 incorporates essential features such as agentic tool use, and its multilingual capabilities are extended to support Spanish in addition to English and Korean. The EXAONE 4.0 model series consists of two sizes: a mid-size 32B model optimized for high performance, and a small-size 1.2B model designed for on-device applications. The EXAONE 4.0 demonstrates superior performance compared to open-weight models in its class and remains competitive even against frontier-class models. The models are publicly available for research purposes and can be easily downloaded via `https://huggingface.co/LGAI-EXAONE`.

1 Introduction

As part of LG AI Research's EXAONE foundation model series, the EXAONE language models have been developed to support diverse real-world applications through strong instruction-following and reasoning capabilities.

The previous version, EXAONE 3.5 [31], focused on real-world usability by strengthening comprehensive instruction-following abilities, while EXAONE Deep [32] emphasized reasoning performance, particularly in mathematical and coding domains.

With the upcoming era of agentic AI in mind, EXAONE 4.0 introduces agentic tool use—a core capability for this paradigm—and further advances reasoning abilities.

In terms of tool use, the model is developed to enable the integration of various external tools to develop agents or applications. Regarding reasoning performance, the capabilities of EXAONE 4.0 have been improved by leveraging the validated methodologies developed in EXAONE Deep. Notably, EXAONE 4.0 unifies both NON-REASONING mode—enabling rapid thinking and responses—and REASONING mode—designed for deep thinking and more accurate answers—into a single model, allowing users to experience both modes within one model.

Compared to previous versions of EXAONE, the number of tokens used during pretraining is significantly increased to bolster world knowledge. To further enhancement of expert knowledge, curating training data from specialized domains such as STEM (Science, Technology, Engineering, and Mathematics) fields plays important role on downstream tasks. Furthermore, the extension of maximum context length of the model to 128K tokens enables handling of various tasks based on significantly longer contexts, thereby improving usability. One notable challenge in processing long contexts is the computational burden of attention calculations. To mitigate this, a hybrid architecture combining global attention and local attention is adopted. This approach minimizes performance degradation while reducing computational costs during training and inference.

Moreover, the EXAONE 4.0 officially add Spanish to its supported languages, expanding beyond its previous bilingual support for English and Korean. The development of Spanish language support was designed to minimize any negative impact on English and Korean performance while maintaining the same tokenizer and vocabulary as the previous EXAONE 3.5 and Deep models.

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

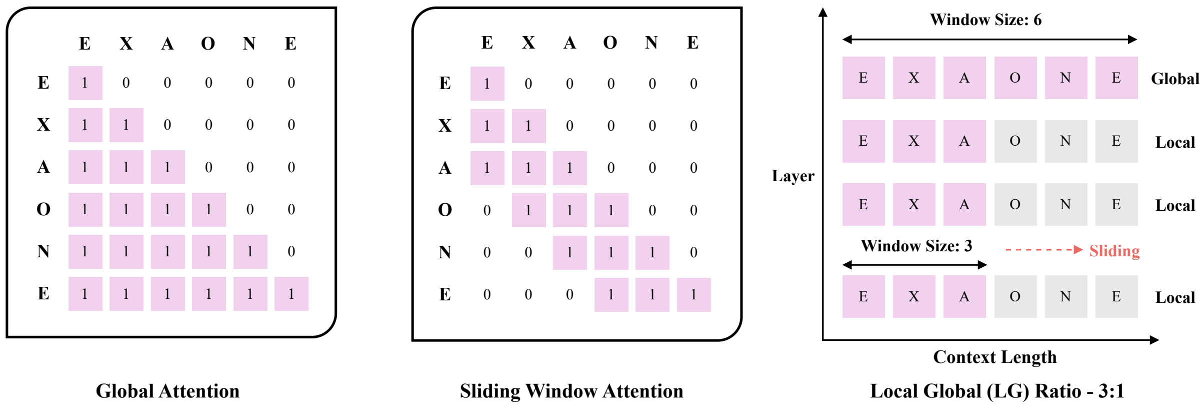

Figure 1: Visualization of the hybrid attention mechanism when the window size for local attention (sliding window attention) is set to 3. This figure illustrates how context tokens are processed across layers under the hybrid attention mechanism, highlighting the interaction between local and global attention.

EXAONE 4.0 particularly excels in areas focused on world knowledge and reasoning, especially in mathematical and coding domains. Despite integrating NON-REASONING mode and REASONING mode, it secures competitive performance in instruction following. The model also shows commendable performance in long context tasks, particularly excelling in document QA (Question Answering) and RAG (Retrieval Augmented Generation) tasks frequently used by real users. Regarding tool use, it achieves a level comparable to competing models, marking the beginning of the fundamental capabilities essential for the upcoming agentic AI era. Additionally, our supported languages now include English, Korean, and the newly supported Spanish. The EXAONE 4.0 model demonstrates competitive performance in both Korean and Spanish across a diverse range of tasks, including expert-level knowledge and mathematical reasoning.

2 Modeling

2.1 Model Configurations

The EXAONE 4.0 model retains a similar structural framework to the EXAONE 3.5 model, but incorporates several key differences in its architecture. Notably, we modify the approach to the attention mechanism. In the previous EXAONE 3.5 model, every layer utilized global attention, whereas the EXAONE 4.0 model employs a hybrid attention mechanism that combines local attention (sliding window attention type) with global attention in a 3:1 ratio, as illustrated in Figure 1.

Contrary to past findings that models using global attention across all layers performed better, recent studies [14, 15, 36] suggested that utilizing a larger window size (e.g., from 512 to 1,024 or 4,096) and applying global attention to only a minority of layers can still achieve excellent long-context performance. Additionally, it reported that incorporating small amounts of global attention periodically in combinations with heterogeneous structures like Mamba [34, 39, 43] helped maintain the ability to understand global context.

In designing the EXAONE 4.0 model, a sliding window size of 4K is selected to minimize any adverse effects on short-context performance. Furthermore, the model does not employ Rotary Position embedding [63] for global attention, ensuring that the model does not develop biases towards length and can maintain a global view. For the design of the local attention mechanism, we do not employ the chunked attention strategy. Instead, we adopt sliding window attention, a well-established form of sparse attention that offers strong theoretical stability. Unlike chunked attention, sliding window attention benefits from wide support in open-source frameworks, ensuring robust implementation and ease of integration. To prevent performance degradation in short-context areas during long context fine-tuning, the EXAONE 4.0 model employs a careful data selection methodology and a progressive training recipe, effectively balancing efficiency and performance.

Another significant change in the EXAONE 4.0 model is the repositioning of layer normalization (LayerNorm), as shown in Figure 2. According to recent studies [53], some layers that do not significantly impact model performance are found mainly in deep layers. This issue is attributed to the Pre-LN transformer architecture [61], which enhances stability but leads to exponentially increasing variance in outputs as model depth increases. A simple operation to control variance by providing more scaling to outputs as layers deepen was proposed, but we find that the QK-Reorder-LN

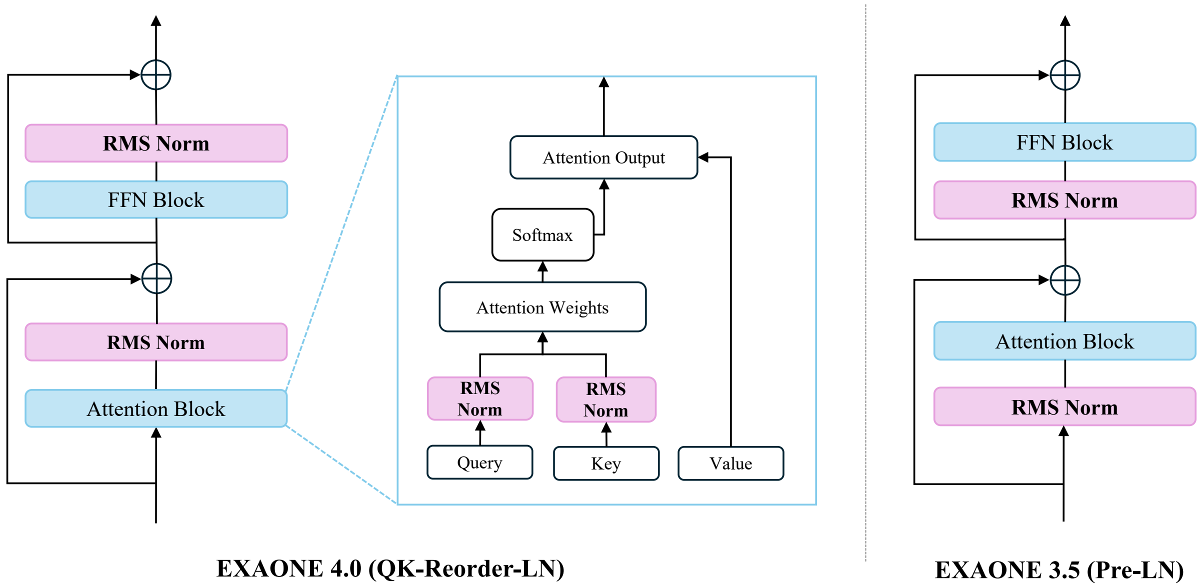

Figure 2: Visualization of repositioning layer normalization. The LayerNorm is applied after input queries and keys, and it is performed after attention output again. The type of normalization is RMSNorm.

method [42, 56], which applies LayerNorm after input queries and keys and performs LayerNorm after attention output, yields better performance on downstream tasks despite consuming more computation. The normalization type RMSNorm, which was applied since EXAONE 3.0, is retained in EXAONE 4.0.

Finally, the EXAONE 4.0 model series consists of two configurations: 32B and 1.2B. These models share the same vocabulary, which consists primarily of Korean and English tokens in roughly equal proportions, along with a tiny number of multilingual tokens, as detailed in Table 1.

Model size	32B	1.2B
*d*_model	5,120	2,048
Number of layers	64	30
Normalization	QK-Reorder-LN	QK-Reorder-LN
Non-linearity	SwiGLU [50]	SwiGLU
Feedforward dimension	27,392	4,096
Attention type	Hybrid	Global
Head type	GQA [4]	GQA
Number of heads	40	32
Number of KV heads	8	8
Head size	128	64
Max sequence length	131,072	65,536
RoPE theta [52]	1,000,000	1,000,000
Tokenizer	BBPE [58]	BBPE
Vocab size	102,400	102,400
Tied word embedding	False	True
Knowledge cut-off	Nov. 2024	Nov. 2024

Table 1: Configurations of EXAONE 4.0 language models. Key differences from previous versions include a hybrid attention mechanism and modified normalization.

Model size	32B	1.2B
Size of pretraining data (tokens)	14T	12T
Amount of computation (FLOPs)	2.69×10^{24}	8.65×10^{22}

Table 2: Pretraining data size and computational resources used for EXAONE 4.0 language models. EXAONE 4.0 utilizes nearly twice the data of its predecessor, EXAONE 3.5.

2.2 Pre-training

The amount of data and computational resources used for pretraining in the EXAONE 4.0 models are summarized in Table 2. For the EXAONE 3.5 32B model, 6.5 trillion tokens are used for pretraining. In comparison, the EXAONE 4.0 32B model doubles this amount, utilizing 14 trillion tokens for pretraining. This increase in data is specifically aimed at enhancing the model's world knowledge. As will be discussed later, this approach yields noticeable improvements in benchmarks that rely on knowledge, such as MMLU-Redux [13], where the use of more extensive training data has a demonstrable impact on performance.

Furthermore, as recent studies showed that reasoning performance was significantly influenced by the cognitive behavior [12] acquired from documents seen during pretraining, we perform rigorous data curation during pretraining to enhance post-training performance.

2.3 Context Length Extension

In the EXAONE 4.0 model, the maximum context length is extended to 128K tokens. To achieve this, we undertake a two-stage context length extension process. Initially, a model pretrained with a context length of 4K tokens is firstly extended to 32K tokens. Subsequently, it is further extended to 128K tokens.

The long-context fine-tuning process is meticulously executed, with the Needle In A Haystack (NIAH) test [16] at each stage to ensure thorough validation of the model's performance. This iterative refinement continues until comprehensive optimization is achieved and the "green light" signal is consistently observed across all segments, signifying the successful extension of the context length to 128K tokens without compromising the model's overall performance.

For the 1.2B model, the context length is extended up to 64K tokens, which is approximately twice as long as the typical maximum length of 32K tokens supported by most models in the 1B-parameter range.

2.4 Post-training

In EXAONE 4.0, multiple stages of training is undertaken to enable the model to respond to a variety of user instructions and integrate NON-REASONING and REASONING models effectively. The training process is primarily organized into three stages: supervised fine-tuning (SFT), reasoning reinforcement learning (RL), and preference learning to integrate NON-REASONING and REASONING modes as illustrated in Figure 3.

A significant feature of the post-training phase is the large-scale expansion of the SFT data to enhance performance in an efficient manner. To improve reasoning capabilities, RL is employed. Additionally, a hybrid reward mechanism is used in a two-stage preference learning process to seamlessly integrate NON-REASONING and REASONING modes.

2.4.1 Large-scale Supervised Fine-tuning

The composition of the SFT dataset is divided into non-reasoning and reasoning data. Furthermore, it can be classified into five distinct domains: World Knowledge, Math/Code/Logic, Agentic Tool Use, Long Context, and Multilinguality. Data collection and generation strategies were differentiated for each purpose and domain, and the detailed methodologies are described below.

World Knowledge For the world knowledge domain, which encompasses a wide range of fields and levels of difficulty, it is essential to enable the distillation of extensive knowledge. Therefore, we filtered problems collected from web sources based on their educational value, prioritizing the use of high-quality data. Among these, we also sample specialized and high-difficulty data to utilize in training for REASONING mode.

Math, Code, Logic For the Math, Code, and Logic tasks, the number of unique problems is relatively limited compared to their importance. This is primarily because establishing accurate ground truth is not only essential but

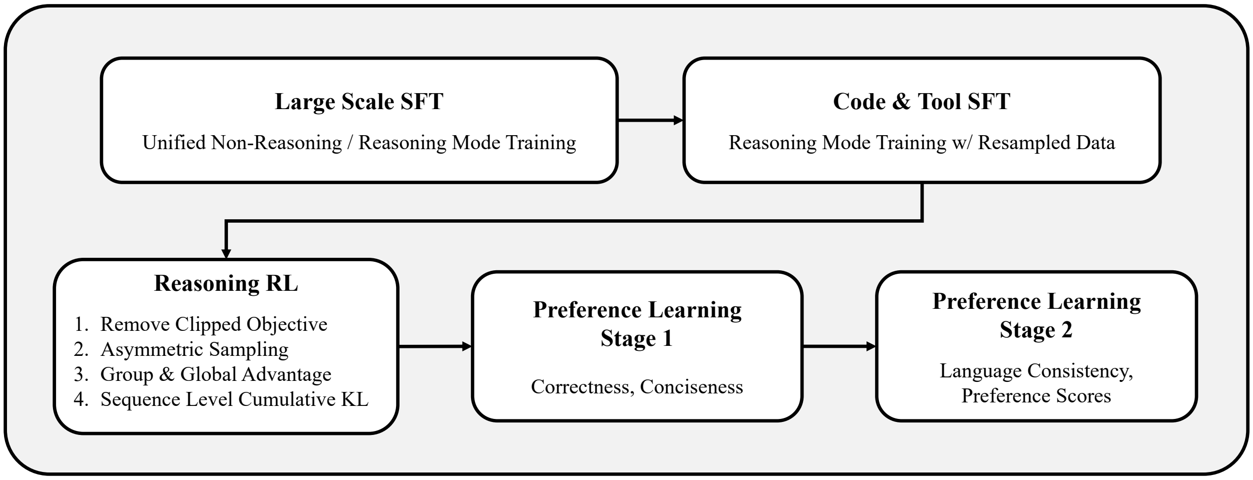

Figure 3: The post-training pipeline of the EXAONE 4.0. The pipeline consists of five stages, which include supervised fine-tuning (SFT), reinforcement learning (RL), and preference learning.

also difficult in these domains, thereby limiting our ability to construct as many high-quality problems as desired. Consequently, rather than create unverifiable problems, we train on diverse responses for queries with verifiable answers, and observe that generating multiple responses per unique query is as effective as increasing the diversity or number of unique queries themselves. Furthermore, in the REASONING mode, responses for Math and Code domains tend to be longer, which increases the risk of degeneration and language inconsistency; thus, careful filtering is applied. Additionally, for the Code domain, we extend our data collection beyond problem-solving to include a software engineering dataset focused on full stack development, created from code corpora.

Long Context We construct a long-context SFT dataset from web corpora, focusing on tasks that require comprehensive understanding of extended inputs. To train models to identify and reason over dispersed information, we systematically vary both the context length and the location of key content. The dataset also includes instruction-following queries for long-form generation, allowing models to produce coherent and well-structured long outputs. For Korean, we curate long-context data by refining documents such as legal, administrative, and technical texts. These documents are then restructured to accommodate a diverse range of long-context input formats, ensuring variation in structure and content scope.

Agentic Tool Use To enhance the model's capability for agentic tool use, we construct datasets focused on both single-turn and multi-turn tasks, leveraging diverse tool lists. Rather than merely creating datasets for single tool calls, we emphasize the construction of more complex, long-horizon tool-calling data. Accordingly, we develop user-agent conversations that incorporate user interaction, execution feedback from the environment, and iterative reasoning, ultimately guiding the agent to achieve the user's desired goal. These datasets are organized in multi-step and multi-turn formats to better support the learning of agentic tool use.

Multilinguality To support both Korean and Spanish, we construct datasets that not only target cultural and historical knowledge specific to each language, but also enable the model to engage in fluent, natural conversations with users. We create new instructions in both languages and additionally leveraged translations of selected existing samples as queries. For Korean, in particular, we curate data to address topics relevant to local education and industry experts, ensuring that the model is well-equipped to handle domain-specific queries from Korean users.

Unified Mode Training In the combined dataset, the NON-REASONING data primarily consists of diverse tasks, while the REASONING data is centered on Math and Code domains. Rather than fine-tuning the two modes sequentially, we combine both modes and train them together. The ratio between the two modes is determined by the amount of Reasoning mode data. If the token ratio of REASONING mode is too high, we observe that the model tends to behave as if it is in REASONING mode even when NON-REASONING mode is enabled. Through ablation studies, we set the token ratio of REASONING to NON-REASONING data to 1.5:1.

After unified NON-REASONING/REASONING mode fine-tuning, to address domain imbalance, we perform a second round of training using high-quality REASONING data from the Code and Tool Use domains, reusing these samples to further enhance the performance.

2.4.2 Reasoning Reinforcement Learning

To enhance the model's reasoning capabilities, we conduct online reinforcement learning (RL) following supervised fine-tuning (SFT). Previous studies demonstrated that combining the GRPO (Group Relative Policy Optimization) algorithm [49] with verifiable rewards [28] can effectively improve model performance. To address the limitations of existing GRPO, we propose a new algorithm, named AGAPO (Asymmetric Sampling and Global Advantage Policy Optimization).

Our training dataset encompasses curated data across four categories: mathematics, code, science, and instruction following. To focus training on more informative data samples, we perform accuracy-based filtering by generating eight responses from the SFT model and excluding samples where all eight responses are correct, a pre-filtering step that removes problems that are easy for the model to avoid inefficient training.

The reward function used in RL is tailored for each category. For the mathematics category, a rule-based verifier is used to determine correctness. In the code category, a response is considered correct if its final code block passes all associated test cases. For the science category, a rule-based verifier is first applied; if a response is deemed incorrect, an LLM-judge then performs a more flexible verification. Finally, for the instruction following category, a reward of 1 is assigned if all constraints are satisfied, and 0 otherwise.

For the algorithm design, AGAPO comprehensively improves upon existing methods. Its main features are as follows:

- **Remove Clipped Objective.** Previous research has questioned the necessity of PPO(Proximal Policy Optimization) [48]'s clip loss [3] and shown that this clipped objective can degrade performance [40] by preventing crucial, low-probability tokens from contributing to gradient updates. These tokens are often associated with reflective behaviors that serve as forks in the reasoning path. AGAPO removes the clipping from PPO and instead uses a standard policy gradient loss. This approach is designed to prevent the dropping of these exploratory tokens, allowing for more substantial policy updates while maintaining training stability.
- **Asymmetric Sampling.** Previous works filter out samples where all responses were either correct or incorrect [17, 67] because they result in a zero advantage for GRPO. However, as recent work has shown the effectiveness of Negative Sample Reinforcement [70], AGAPO utilizes an asymmetric sampling method that does not discard samples where all responses are incorrect, thereby including a higher proportion of negative feedback. For these all-incorrect samples, a small negative reward is assigned through the advantage calculation, allowing them to be used to guide the model away from erroneous reasoning paths.
- **Group&Global Advantages.** GRPO advantage method does not account for the distribution of the entire batch, which makes it difficult to assign appropriate negative rewards to groups of all-incorrect samples. To improve this, AGAPO calculates the advantage in two stages: group and global. First, at the group level, the advantage is computed using the Leave-One-Out(LOO) method [3] within a response group based on verifiable reward accuracy. Next, normalization is performed across the entire batch (global) to calculate a final advantage that considers the full batch distribution.
- **Sequence Level Cumulative KL.** To enhance performance while preserving capabilities learned during the SFT stage, a KL penalty is applied. We adopt the sequence-level cumulative KL [54], as proposed in prior research, to ensure the model receives an appropriate gradient during training.

Objective The AGAPO objective is defined for a question q sampled from the training distribution $P(Q)$. For each question, the current policy $\pi_\theta(\cdot \mid q)$ generates a *group* of G candidate responses, denoted as $O = \{o_1, \ldots, o_G\}$. Each response o_i is assigned a verifiable reward $r_i \in [0, 1]$. The objective function maximizes the following:

$$\mathcal{J}_{\text{AGAPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_\theta(O|q)}\Big[\frac{1}{G}\sum_{i=1}^{G}\Big(A_{\text{global},i}\ \log\pi_\theta(o_i \mid q) - \beta\, D_{\text{KL}}\big(\pi_\theta, \pi_{\text{ref}}\big)\Big)\Big]. \tag{1}$$

The global advantage $A_{\text{global},i}$ in the objective is calculated in two stages. First, a leave-one-out (LOO) advantage is computed within each group. This advantage is then normalized across the entire mini-batch of size $K = B \times G$ to yield the final global advantage:

$$A_{\text{loo},i} = r_i - \frac{1}{G-1}\sum_{j\neq i} r_j, \quad A_{\text{global},i} = \frac{A_{\text{loo},i} - \text{mean}\big(\{A_{\text{loo},k}\}_k\big)}{\text{std}\big(\{A_{\text{loo},k}\}_k\big)}. \tag{2}$$

2.4.3 Preference Learning with Hybrid Reward

In the RL stage, we aim to enhance accuracy through verifiable rewards, and do not use human preference. In addition, since the model is specialized for reasoning tasks, we observe a decline in performance in other types of tasks. To overcome these limitations, we introduce an additional preference learning phase.

Preference learning is conducted by directly learning human preferences from chosen and rejected data pairs, akin to the Direct Policy Optimization (DPO) framework [46]. We employ SimPER [60], among various reference-free preference optimization methods for this learning process. The dataset for preference learning is constructed based on on-policy responses [28, 38] generated by the model after completing the RL phase. For each query, we generate 4 to 16 responses per task, and select chosen and rejected responses based on a hybrid reward combining verifiable reward, preference reward, language consistency reward, and conciseness reward, tailored per task.

The training is conducted separately for two stages. In the first stage, we focus on increasing token efficiency by reducing the generation length while maintaining the performance of the reasoning mode. Therefore, for reasoning-related verifiable training data, we combine the verifiable reward with a conciseness reward to select the shortest response among the correct answers as the chosen option. In the second stage, we employ a combination of preference reward and language consistency reward for human alignment. For the REASONING Mode data, preference labeling is performed only on the final answer after the reasoning process is complete. Furthermore, to ensure stability during the second stage of training, a portion of the data from the first stage is sampled and reused.

2.5 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 [30] and the LG AI Ethics Principles [29].

3 Evaluation

3.1 Benchmarks

We evaluate EXAONE 4.0 on a diverse set of benchmarks spanning 6 categories: World Knowledge, Math/Coding, Instruction Following, Long Context, Agentic Tool Use, and Multilinguality.

- **World Knowledge** We select benchmarks to evaluate the extent of our model's world knowledge, including MMLU-REDUX [13] and MMLU-PRO [59], a refined and extended version of MMLU [19]. Additionally, we utilize GPQA-DIAMOND [47] to assess the expert-level knowledge in Biology, Physics, and Chemistry.
- **Math/Coding** Challenging benchmarks in Math and Coding categories are adopted to evaluate the test-time computational capability of EXAONE 4.0. For Math, we utilize two math Olympiad competitions: AIME 2025 [37] and HMMT FEB 2025 [6]. For Coding, LIVECODEBENCH V5 and V6 [24] are chosen.
- **Instruction Following** To evaluate how well our models understand and align with users' instructions, we select IFEVAL [69] and MULTI-IF [18], the latter being an extension of IFEVAL to support multi-turn and multilingual scenarios. We use only the English subset of MULTI-IF to focus on assessing the multi-turn instruction-following ability on English.
- **Long Context** To evaluate the model's ability to understand and solve tasks requiring long-context comprehension, we adopt three representative benchmarks: HELMET [66], RULER [22], and LONGBENCH [5]. These benchmarks collectively cover both synthetic tasks and real-world scenarios. To maintain a coherent evaluation focused on core long-context capabilities, we exclude the *LongCite* task from HELMET (see Appendix D for further details).
- **Agentic Tool Use** With the advancement of LLM-based agents, numerous benchmarks have emerged to evaluate their tool-use capabilities, among which we focus on the two most widely adopted: BFCL-V3 [44] and TAU-BENCH [65]. BFCL-V3 evaluates various aspects of function-calling abilities. TAU-BENCH assesses tool calling performance through simulated conversations between a user LLM. We utilize *gpt-4.1-2025-04-14* model as the user role.
- **Multilinguality** Beyond English, we evaluate our models on two additional languages: Korean and Spanish. For Korean, we use KMMLU-PRO[1] for measuring practical applicability on professional knowledge and KMMLU-

[1] https://huggingface.co/datasets/LGAI-EXAONE/KMMLU-Pro

	Mid-size				Frontier	
	EXAONE 4.0 32B (Reasoning)	Phi 4 reasoning-plus	Magistral Small-2506	Qwen 3 32B (Reasoning)	Qwen 3 235B (Reasoning)	DeepSeek R1 -0528
Type	Hybrid	Reasoning	Reasoning	Hybrid	Hybrid	Reasoning
# Total Params	32.0 B	14.7 B	23.6 B	32.8 B	235 B	671 B
World Knowledge						
MMLU-Redux	92.3	90.8	86.8	90.9*	92.7*	93.4*
MMLU-Pro	81.8	76.0*	73.4	80.0	83.0	85.0*
GPQA-Diamond	75.4	68.9*	68.2*	68.4*	71.1*	81.0*
Math / Coding						
AIME 2025	85.3	78.0*	62.8*	72.9*	81.5*	87.5*
HMMT Feb 2025	72.9	53.6*	43.5	50.4	62.5*	79.4*
LiveCodeBench v5	72.6	51.7	55.8*	65.7*	70.7*	75.2*
LiveCodeBench v6	66.7	47.1	47.4*	60.1	58.9*	70.3*
Instruction Following						
IFEval	83.7	84.9*	37.9	85.0*	83.4*	80.8
Multi-IF (EN)	73.5	56.1	27.4	73.4	73.4	72.0
Agentic Tool Use						
BFCL-v3	63.9	N/A	40.4	70.3*	70.8*	64.7*
Tau-Bench (Airline)	51.5	N/A	38.5	34.5	37.5	53.5*
Tau-Bench (Retail)	62.8	N/A	10.2	55.2	58.3	63.9*
Multilinguality						
KMMLU-Pro (KO)	67.7	55.8	51.5	61.4	68.1	71.7
KMMLU-Redux (KO)	72.7	62.7	54.6	67.5	74.5	77.0
KSM (KO)	87.6	79.8	71.9	82.8	86.2	86.7
MMMLU (ES)	85.6	84.3	68.9	82.8*	86.7*	88.2
MATH500 (ES)	95.8	94.2	83.5	94.3	95.1	96.0

Table 3: The main evaluation results of EXAONE 4.0 32B Reasoning mode. Missing entries (`N/A`, Not Applicable) indicate that the corresponding model does not support the given input length or task. Asterisk (*) indicates that the scores are from each baseline model's official technical report, blog or leaderboard.

Redux[2] for assessing real-world expert knowledge instead of KMMLU [51] to ensure benchmark reliability. KMMLU have been reported dataset error and contamination issue between pre-training corpora and task dataset. In addition, we employ Korean School Math (KSM) subset of HRM8K [26] to evaluate a wide range of Korean mathematical knowledge from high-school to Olympiad level. To evaluate the models' ability to handle long-context Korean inputs, we also include an in-house benchmark, Ko-LongBench (Please refer to Appendix D.4 for details). For Spanish, we adopt the translated version of existing benchmarks. To be specific, we use MMMLU (ES) [3] and MATH500 [35] (ES) [4]. Furthermore, we assess translation ability using WMT24++ [10], a widely-used translation benchmark. We consider only EN and ES pair, and utilize LLM-as-a-judge [5] to score the translation quality.

3.2 Baselines

To evaluate the performance of language models from various perspectives, recently released open-weight models are selected as baseline models. These baseline models include not only models of similar sizes but also frontier-level models exceeding 100B parameters, which exhibits superior performance. These models can be divided into three types: (1) Non-Reasoning models, which generate their responses in CoT (Chain-of-Thought) style, (2) Reasoning models,

[2] `https://huggingface.co/datasets/LGAI-EXAONE/KMMLU-Redux`

[3] `https://huggingface.co/datasets/openai/MMMLU`

[4] `https://huggingface.co/datasets/bezir/MATH-500-multilingual`

[5] *gpt-4.1-2025-04-14* is used for the judge model. We follow reference-based direct assessment method used in WMT24++ [10]. The exact prompt used for judge is in Appendix D.5.

	Mid-size					Frontier		
	EXAONE 4.0 32B (Non-reasoning)	Phi 4	Mistral Small-2506	Gemma 3 27B	Qwen 3 32B (Non-reasoning)	Qwen 3 235B (Non-reasoning)	Llama 4 Maverick	DeepSeek V3 -0324
Type	Hybrid	Non-Reasoning	Non-Reasoning	Non-Reasoning	Hybrid	Hybrid	Non-Reasoning	Non-Reasoning
# Total Params	32.0 B	14.7 B	24.0 B	27.4B	32.8 B	235 B	402 B	671 B
World Knowledge								
MMLU-Redux	89.8	88.3	85.9	85.0	85.7*	89.2*	92.3	92.3
MMLU-Pro	77.6	70.4*	69.1*	67.5*	74.4	77.4	80.5*	81.2*
GPQA-Diamond	63.7	56.1*	46.1*	42.4*	54.6*	62.9*	69.8*	68.4*
Math / Coding								
AIME 2025	35.9	17.8	30.2	23.8	20.2*	24.7*	18.0	50.0*
HMMT Feb 2025	21.8	4.0	16.9	10.3	9.8	11.9	7.3	29.2*
LiveCodeBench v5	43.3	24.6	25.8	27.5	31.3*	35.3*	43.4*	46.7
LiveCodeBench v6	43.1	27.4	26.9	29.7	28.0	31.4	32.7	44.0
Instruction Following								
IFEval	84.8	63.0*	77.8	82.6	83.2*	83.2*	85.4	81.2
Multi-IF (en)	71.6	47.7	63.2	72.1	71.9	72.5	77.9	68.3
Long Context								
HELMET	58.3	N/A	61.9	58.3*	54.5	63.3	13.7	N/A
RULER	88.2	N/A	71.8	66.0*	85.6*	90.6*	2.9	N/A
LongBench v1	48.1	N/A	51.5	51.5	44.2	45.3	34.7	N/A
Agentic Tool Use								
BFCL-v3	65.2	N/A	57.7*	N/A	63.0*	68.0*	52.9*	63.8*
Tau-Bench (Airline)	25.5	N/A	36.1	N/A	16.0	27.0	38.0	40.5
Tau-Bench (Retail)	55.9	N/A	35.5	N/A	47.6	56.5	6.5	68.5
Multilinguality								
KMMLU-Pro (ko)	60.0	44.8	51.0	50.7	58.3	64.4	68.8	67.3
KMMLU-Redux (ko)	64.8	50.1	53.6	53.3	64.4	71.7	76.9	72.2
KSM (ko)	59.8	29.1	35.5	36.1	41.3	46.6	40.6	63.5
Ko-LongBench (ko)	76.9	N/A	55.4	72.0	73.9	74.6	65.6	N/A
MMMLU (es)	80.6	81.2	78.4	78.7	82.1*	83.7*	86.9	86.7
MATH500 (es)	87.3	78.2	83.4	86.8	84.7	87.2	78.7	89.2
WMT24++ (es)	90.7	89.3	92.2	93.1	91.4	92.9	92.7	94.3

Table 4: The main evaluation results of EXAONE 4.0 32B Non-reasoning mode. Missing entries (N/A, Not Applicable) indicate that the corresponding model does not support the given input length or task. Asterisk (*) indicates that the scores are from each baseline model's official technical report, blog or leaderboard.

which generate in long CoT style, and (3) Hybrid model, which generate in either CoT or long CoT style depending on the mode. Detailed information about the models is presented in the Appendix C.

3.3 Experimental Setup

Hyperparameters We sample n different responses for each problem in benchmarks with limited examples to ensure evaluation stability. Specifically, we sample $n = 8$ responses for GPQA-Diamond, $n = 32$ for AIME 2025 and HMMT Feb 2025, and $n = 4$ for LiveCodeBench v5/6, Tau-Bench and MATH500 (es). The accuracy is averaged over the n samples. In Reasoning mode, we set temperature to 0.6, top-p [21] to 0.95, and apply a presence penalty of 1.5 only for our 32B model. In contrast, for Non-Reasoning mode, greedy decoding is used for a single ($n = 1$) generated response, while the same sampling settings as Reasoning mode (except with a presence penalty of 0.0) are used when generating $n > 1$ responses. We generate a maximum of 64K tokens for AIME 2025, HMMT Feb 2025, LiveCodeBench v5/6, and KSM benchmarks, while 32K for other benchmarks.

Long-Context Evaluation of Small-Size models In evaluating long-context performance of Small-Size NON-REASONING models, we extend the context lengths of Qwen3 1.7B and Qwen3 0.6B beyond their 32K token limit

	EXAONE 4.0 1.2B (REASONING)	EXAONE Deep 2.4B	Qwen 3 0.6B (REASONING)	Qwen 3 1.7B (REASONING)	SmolLM 3 3B (REASONING)
	SMALL-SIZE				
Type	Hybrid	Reasoning	Hybrid	Hybrid	Hybrid
# Total Params	1.28 B	2.41 B	596 M	1.72 B	3.08 B
World Knowledge					
MMLU-REDUX	71.5	68.9	55.6*	73.9*	74.8
MMLU-PRO	59.3	56.4*	38.3	57.7	57.8
GPQA-DIAMOND	52.0	54.3*	27.9*	40.1*	41.7*
Math / Coding					
AIME 2025	45.2	47.9*	15.1*	36.8*	36.7*
HMMT FEB 2025	34.0	27.3	7.0	21.8	26.0
LIVECODEBENCH V5	44.6	47.2	12.3*	33.2*	27.6
LIVECODEBENCH V6	45.3	43.1	16.4	29.9	29.1
Instruction Following					
IFEVAL	67.8	71.0	59.2*	72.5*	71.2*
MULTI-IF (EN)	53.9	54.5	37.5	53.5	47.5
Agentic Tool Use					
BFCL-V3	52.9	N/A	46.4*	56.6*	37.1
TAU-BENCH (Airline)	20.5	N/A	22.0	31.0	37.0
TAU-BENCH (Retail)	28.1	N/A	3.3	6.5	5.4
Multilinguality					
KMMLU-PRO (KO)	42.7	24.6	21.6	38.3	30.5
KMMLU-REDUX (KO)	46.9	25.0	24.5	38.0	33.7
KSM (KO)	60.6	60.9	22.8	52.9	49.7
MMMLU (ES)	62.4	51.4	48.8*	64.5*	64.7
MATH500 (ES)	88.8	84.5	70.6	87.9	87.5

Table 5: The main evaluation results of EXAONE 4.0 1.2B REASONING mode. Missing entries (`N/A`, Not Applicable) indicate that the corresponding model does not support the given input length or task. Asterisk (*) indicates that the scores are from each baseline model's official technical report, blog or leaderboard.

by applying YaRN [45], enabling inference up to 64K tokens. For reference, evaluation results of models such as Gemma-3-1B, EXAONE-3.5-2.4B-Instruct, Qwen3 1.7B, and Qwen3 0.6B at context lengths up to 32K tokens are provided in the Appendix D.

Baselines Reproduction For baseline models, we borrow scores reported in each model's official technical report, blog, or leaderboard[6] if available. If not, we reproduce the results in our evaluation environment, following the recommended settings when they are explicitly stated[7].

3.4 Experimental Results

Table 3, 4, 5, and 6 present the benchmark performances of our EXAONE 4.0 models in both REASONING and NON-REASONING modes. The key results are summarized below:

Superiority in Math/Coding domains EXAONE 4.0 models demonstrate extraordinary performance in Math/Coding benchmarks. Specifically, EXAONE 4.0 32B model outperforms Qwen3 235B in both REASONING and NON-REASONING modes across all Math/Coding benchmarks. At the same time, EXAONE 4.0 1.2B model surpasses all baselines, except for EXAONE Deep 2.4B in REASONING mode.

[6] We refer to `https://github.com/LiveCodeBench/submissions` for LIVECODEBENCH, `https://matharena.ai/` for HMMT FEB 2025, and `https://gorilla.cs.berkeley.edu/leaderboard.html` for BFCL-V3.

[7] For example, the Qwen3 series explicitly specifies recommended decoding parameters in its Hugging Face repository.

	Small-size				
	EXAONE 4.0 1.2B (Non-reasoning)	Qwen 3 0.6B (Non-reasoning)	Gemma 3 1B	Qwen 3 1.7B (Non-reasoning)	SmolLM 3 3B (Non-reasoning)
Type	Hybrid	Hybrid	Non-Reasoning	Hybrid	Hybrid
# Total Params	1.28 B	596 M	1.00 B	1.72 B	3.08 B
World Knowledge					
MMLU-Redux	66.9	44.6*	40.9	63.4*	65.0
MMLU-Pro	52.0	26.6	14.7*	43.7	43.6
GPQA-Diamond	40.1	22.9*	19.2*	28.6*	35.7*
Math / Coding					
AIME 2025	23.5	2.6*	2.1	9.8*	9.3*
HMMT Feb 2025	13.0	1.0	1.5	5.1	4.7
LiveCodeBench v5	26.4	3.6*	1.8	11.6*	11.4
LiveCodeBench v6	30.1	6.9	2.3	16.6	20.6
Instruction Following					
IFEval	74.7	54.5*	80.2*	68.2*	76.7*
Multi-IF (en)	62.1	37.5	32.5	51.0	51.9
Long Context					
HELMET	41.2	21.1	N/A	33.8	38.6
RULER	77.4	55.1	N/A	65.9	66.3
LongBench v1	36.9	32.4	N/A	41.9	39.9
Agentic Tool Use					
BFCL-v3	55.7	44.1*	N/A	52.2*	47.3
Tau-Bench (Airline)	10.0	31.5	N/A	13.5	38.0
Tau-Bench (Retail)	21.7	5.7	N/A	4.6	6.7
Multilinguality					
KMMLU-Pro (ko)	37.5	24.6	9.7	29.5	27.6
KMMLU-Redux (ko)	40.4	22.8	19.4	29.8	26.4
KSM (ko)	26.3	0.1	22.8	16.3	16.1
Ko-LongBench (ko)	69.8	16.4	N/A	57.1	15.7
MMMLU (es)	54.6	39.5*	35.9	54.3*	55.1
MATH500 (es)	71.2	38.5	41.2	66.0	62.4
WMT24++ (es)	65.9	58.2	76.9	76.7	84.0

Table 6: The main evaluation results of EXAONE 4.0 1.2B Non-reasoning mode. Missing entries (N/A, Not Applicable) indicate that the corresponding model does not support the given input length or task. Asterisk (*) indicates that the scores are from each baseline model's official technical report, blog, or leaderboard.

Competitive Performance in Tool Use Scenarios EXAONE 4.0 32B model shows competitive performance in tool use compared to baseline models. For example, in Reasoning mode, it demonstrates similar performance to R1-0528 in Tau-Bench, and achieves comparable BFCL-v3 results with Qwen 3 235B in Non-Reasoning mode. This is noteworthy considering both baselines are much larger than ours. EXAONE 4.0 1.2B model, despite its small size, achieves the highest performance on Tau-Bench (Retail) compared to the baselines.

World Knowledge and GPQA Both our models excel in benchmarks in the World Knowledge category. Despite their relatively smaller size compared to the baselines, they achieve competitive performance. Among the benchmarks, the EXAONE 4.0 models especially demonstrate better performance in GPQA-Diamond. Both EXAONE 4.0 32B and 1.2B models achieve second-highest performance in GPQA-Diamond when Reasoning mode is available.

3.5 Reasoning Budget

We control the number of reasoning tokens and observe how performance varies according to the *reasoning budget*. Specifically, while we set the maximum number of tokens to 64K for benchmarks in Math/Coding categories int the main experiments, in this section we vary the number of tokens used for reasoning from 1K to 64K in this section.

Reasoning Budget	64K	32K	16K	8K	4K	2K	1K
EXAONE 4.0 32B							
AIME 2025	85.3	74.8	44.2	36.8	35.5	35.7	35.6
LiveCodeBench v6	66.7	67.3	53.0	47.6	46.0	45.7	44.0
EXAONE 4.0 1.2B							
AIME 2025	45.2	45.3	37.1	24.6	23.2	22.7	22.3
LiveCodeBench v6	45.3	43.0	40.1	38.3	34.0	33.4	29.3

Table 7: The results of controlling the *reasoning budget* of EXAONE 4.0 models on AIME 2025 and LiveCodeBench v6. The reasoning budget indicates the number of tokens used for reasoning part of the model response. We fix the length of the answer part to 8K.

Similar to [62], when the model's generation reaches the maximum token budget, we stop the generation, append the text *"Considering the limited time by the user, I have to give the solution based on the thinking directly now.\n</think>\n\n"*, and proceed to generate the answer part. We use same number of sampled responses per each query as in the main experiments ($n = 32$ for AIME 2025 and $n = 4$ for LiveCodeBench v6) and average the result over n responses. We fix the length of the answer part to 8K.

The result is presented in Table 7. While a reduced reasoning budget leads to some performance degradation, our EXAONE 4.0 models still demonstrate competitive performance even with a 32K reasoning budgets. Specifically, except for the 32B model on AIME 2025, which shows a 12.3% decrease in performance, the decrease for others is similar or less than 5%, maintaining competitive results compared to baseline models.

4 Limitations

EXAONE 4.0 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 make 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 4.0 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 models do 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 4.0 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 4.0 language models.

5 Deployment

Section B in the Appendix provides license information for using the EXAONE 4.0 models. Understanding the license information is essential for the legal utilization of the language model.

6 Conclusion

In this technical report, we introduce EXAONE 4.0, which integrates NON-REASONING mode and REASONING mode. The key features of EXAONE 4.0 include enhancing the practical usability and reasoning capabilities previously supported in EXAONE 3.5 and EXAONE Deep, consolidating them into a single model, and introducing new functionalities such as agentic tool use and support for Spanish. In terms of performance, EXAONE 4.0 demonstrates superior results compared to models of similar scale and achieves competitive performance even compared to frontier models. As part of our future work, we aim to continuously strengthen usability by gradually expanding the supported languages.

Since the release of EXAONE 3.0, LG AI Research has contributed to the expansion of the research ecosystem by publicly disclosing the model in an open-weight format, and has been continuously improving the model based on user feedback. For any improvement suggestions or business-related inquiries regarding the model, please contact us at 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, Jiyeon Jung, 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 Kyunghoon Bae, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Taewan Hwang, Joonwon Jang, Kijeong Jeon, Gerrard Jeongwon Jo, Euisoon Kim, Hyosang Kim, Jihoon Kim, Youchul Kim, Edward Hwayoung Lee, Gwangho Lee, Honglak Lee, Yongmin Park, Young Min Paik, Youngyong Park, Sanghyun Seo, Sihoon Yang, Sihyuk Yi

B Model License

EXAONE AI Model License Agreement 1.2 - 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 and educational purposes. This includes evaluation, testing, academic research, experimentation, learning, teaching, training 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 and educational 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 any models that compete with the Licensor's 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: Licensor claims no rights in Output. Licensee is solely responsible for the Output and its use.

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 and Derivatives and destroy all copies of the Model and Derivatives 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 Baseline models

The models being compared are categorized into open-weight models: Small-size models under 3B, Mid-size models between 10B and 30B, and Frontier models above 200B. Additionally, the models are divided into three types for performance evaluation, with specific details provided in Table 8.

Category	Model	Parameters	Type	Release Date
Frontier	DeepSeek R1-0528 [8]	671B (MoE)	Reasoning	May 2025
	DeepSeek V3-0324 [9]	671B (MoE)	Non-reasoning	Mar. 2025
	Llama 4 Maverick	402B (MoE)	Non-reasoning	Apr. 2025
	Qwen 3 235B [62]	235B (MoE)	Hybrid	Apr. 2025
Mid-size	Qwen 3 32B [62]	32.8B	Hybrid	Apr. 2025
	EXAONE 4.0 32B	32.0B	Hybrid	Jul. 2025
	Gemma 3 27B [55]	27.4B	Non-reasoning	Mar. 2025
	Mistral-Small-3.2-24B-Instruct-2506	24.0B	Non-reasoning	Jun. 2025
	Magistral-Small-2506 [41]	23.6B	Reasoning	Jun. 2025
	Phi 4 reasoning plus [1]	14.7B	Reasoning	Apr. 2025
	Phi 4 [2]	14.7B	Non-reasoning	Dec. 2024
Small-size	SmolLM 3 3B	3.08B	Hybrid	Jul. 2025
	EXAONE Deep 2.4B [32]	2.41B	Reasoning	Mar. 2025
	Qwen 3 1.7B [62]	1.72B	Hybrid	Apr. 2025
	EXAONE 4.0 1.2B	1.28B	Hybrid	Jul. 2025
	Gemma 3 1B [55]	1.00B	Non-reasoning	Mar. 2025
	Qwen 3 0.6B [62]	596M	Hybrid	Apr. 2025

Table 8: The list of EXAONE 4.0 models and baseline models used for the evaluation along with their parameter size, type, and released date.

D Evaluation Details

D.1 HELMET

We include the HELMET benchmark [66] in our evaluation to systematically assess models' long-context capabilities across both synthetic and real-world tasks. HELMET is designed as a comprehensive suite of diverse, application-centric tasks and addresses key limitations of prior benchmarks, such as inadequate input lengths, over-reliance on retrieval-style setups, and unreliable evaluation metrics. Crucially, it covers a wide spectrum of long-context challenges, including information recall, multi-hop retrieval, in-context generalization, and long-input generation, making it a well-suited benchmark for evaluating models' ability to process and reason over extended sequences in practical settings.

We adopt six of the seven categories from HELMET, *Synthetic Recall*, *Retrieval-Augmented Generation (RAG)*, *Passage Re-ranking*, *In-Context Learning (ICL)*, *Long-document Question Answering (LongQA)*, and *Summarization (Summ)*, to provide a balanced and holistic evaluation of long-context understanding and reasoning abilities.

We formalize our decision to exclude *LongCite* task from HELMET along three lines:

- Scope misalignment: HELMET emphasizes general long-context abilities, such as summarization, question answering, retrieval, and reasoning, whereas *LongCite* centers on sentence-level citation accuracy, which constitutes a distinct attribution task rather than a core comprehension or generative skill.
- Metric incompatibility: The benchmark employs standardized metrics like SubEM and model-based scoring, while *LongCite* introduces specialized citation-precision and F1 measures. Integrating these heterogeneous metrics would compromise the uniformity essential for fair model comparison.
- Benchmark coherence: Including a specialized citation task would divert HELMET from its unified objective of comparing long-context reasoning across models. Such an inclusion would introduce extraneous variability and diminish comparative consistency.

Consequently, omitting *LongCite* ensures that HELMET remains a concise, cohesive benchmark focused solely on evaluating long-context language modeling capabilities across properly aligned tasks. Detailed task-wise scores are reported in Table 9, while Figure 4 illustrates how performance on each task varies across different context lengths.

Context Len.	Model	Total Avg.	Recall	RAG	LongQA	Summ	Rerank	ICL
MID-SIZE								
128K	Mistral-Small-2506	61.93	79.82	63.83	70.17	33.32	53.87	70.56
	Qwen3 235B	63.33	85.23	63.85	70.26	39.08	52.89	68.68
	Qwen3 32B	54.47	74.88	58.20	54.67	36.12	48.17	54.78
	Gemma 3 27B	58.34	82.16	64.63	39.59	34.26	55.06	74.38
	LlaMA-4-Maverick	13.72	30.00	15.20	14.14	5.12	2.36	15.48
	EXAONE 4.0 32B	58.34	94.06	54.75	52.31	25.64	48.78	74.52
SMALL-SIZE								
32K	SmolLM 3B	41.25	75.29	49.75	43.55	18.40	21.92	38.60
	Qwen3 1.7B	35.94	50.07	51.00	35.56	17.47	27.89	33.67
	Qwen3 0.6B	21.85	40.33	30.50	26.28	12.73	5.51	15.73
	Gemma 3 1B	15.49	18.86	32.17	24.70	8.03	2.08	7.13
	EXAONE 3.5 2.4B	41.85	73.35	54.08	31.44	18.87	38.69	34.68
	EXAONE 4.0 1.2B	42.50	73.52	47.75	30.43	15.23	26.80	61.27
64K	SmolLM 3B	38.60	67.81	46.88	41.86	19.58	16.80	38.65
	Qwen3 1.7B†	33.83	44.29	48.63	36.04	18.68	21.80	33.55
	Qwen3 0.6B†	21.10	37.13	28.50	27.47	13.85	4.15	15.50
	Gemma 3 1B	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	EXAONE 3.5 2.4B	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	EXAONE 4.0 1.2B	41.17	69.13	46.25	31.44	15.80	22.33	62.10

Table 9: Comparison of MID-SIZE and SMALL-SIZE models across tasks on the HELMET benchmark. (N/A) indicates that models not supporting specific input lengths are omitted from evaluation. (†) denotes that models are extended to 64K context length using YaRN.

Recall

Model	8K	16K	32K	64K	128K
Mistral-Small-2506	79.4	86.9	85.1	84.8	62.8
Qwen3 235B	90.4	89.0	85.8	84.5	76.4
Qwen3 32B	84.6	80.5	78.4	76.9	53.9
Gemma 3 27B	99.7	96.3	94.3	75.3	45.1
LlaMA-4-Maverick	68.9	25.6	22.8	18.8	14.0
EXAONE 4.0 32B	98.8	96.6	96.9	97.0	81.1
SmolLM 3B	85.6	78.2	62.1	45.4	N/A
Qwen3 1.7B	62.9	52.4	34.9	26.9	N/A
Qwen3 0.6B	50.7	39.6	30.7	27.5	N/A
Gemma 3 1B	29.9	16.4	10.2	N/A	N/A
EXAONE 3.5 2.4B	81.9	81.6	56.5	N/A	N/A
EXAONE 4.0 1.2B	76.6	74.5	69.5	55.9	N/A

RAG

Model	8K	16K	32K	64K	128K
Mistral-Small-2506	70.8	68.8	64.2	66.5	48.9
Qwen3 235B	67.5	68.0	63.8	62.8	57.2
Qwen3 32B	67.5	65.5	59.0	55.0	44.0
Gemma 3 27B	69.9	68.1	65.8	62.8	56.6
LlaMA-4-Maverick	57.0	8.5	4.5	2.5	3.5
EXAONE 4.0 32B	64.2	62.8	53.8	51.2	41.8
SmolLM 3B	55.5	52.0	41.8	38.2	N/A
Qwen3 1.7B	55.0	50.5	47.5	41.5	N/A
Qwen3 0.6B	33.0	31.0	27.5	22.5	N/A
Gemma 3 1B	39.5	32.8	24.2	N/A	N/A
EXAONE 3.5 2.4B	57.5	56.2	48.5	N/A	N/A
EXAONE 4.0 1.2B	49.5	50.0	43.8	41.8	N/A

Rerank

Model	8K	16K	32K	64K	128K
Mistral-Small-2506	85.7	66.6	58.2	40.4	18.5
Qwen3 235B	82.8	68.1	48.9	34.8	29.8
Qwen3 32B	75.9	64.2	48.5	32.1	20.1
Gemma 3 27B	79.8	69.1	57.4	39.0	30.0
LlaMA-4-Maverick	11.8	0.0	0.0	0.0	0.0
EXAONE 4.0 32B	75.4	65.1	48.9	31.6	22.9
SmolLM 3B	39.9	21.3	4.5	1.4	N/A
Qwen3 1.7B	42.6	24.1	16.9	3.5	N/A
Qwen3 0.6B	10.2	3.3	3.0	0.1	N/A
Gemma 3 1B	4.7	1.3	0.2	N/A	N/A
EXAONE 3.5 2.4B	55.2	36.5	24.4	N/A	N/A
EXAONE 4.0 1.2B	37.0	26.9	16.6	8.9	N/A

ICL

Model	8K	16K	32K	64K	128K
Mistral-Small-2506	73.2	74.8	75.8	73.4	55.6
Qwen3 235B	70.8	74.0	74.0	64.2	60.4
Qwen3 32B	52.5	53.7	69.6	56.5	41.5
Gemma 3 27B	70.0	73.7	74.2	76.5	77.4
LlaMA-4-Maverick	69.8	2.2	3.0	0.6	1.8
EXAONE 4.0 32B	71.8	74.8	76.0	76.4	73.6
SmolLM 3B	34.0	39.8	42.0	38.8	N/A
Qwen3 1.7B	36.0	32.2	32.8	33.2	N/A
Qwen3 0.6B	15.4	15.8	16.0	14.8	N/A
Gemma 3 1B	12.4	8.7	0.4	N/A	N/A
EXAONE 3.5 2.4B	28.3	34.9	40.8	N/A	N/A
EXAONE 4.0 1.2B	56.0	60.8	67.0	64.6	N/A

LongQA

Model	8K	16K	32K	64K	128K
Mistral-Small-2506	57.6	62.7	73.8	79.9	76.8
Qwen3 235B	59.8	68.2	72.7	69.9	80.7
Qwen3 32B	42.4	58.6	62.6	57.5	52.3
Gemma 3 27B	30.0	36.9	39.7	44.6	46.8
LlaMA-4-Maverick	49.8	9.8	3.2	3.5	4.4
EXAONE 4.0 32B	43.3	51.3	57.6	52.9	56.5
SmolLM 3B	44.0	43.8	42.9	36.8	N/A
Qwen3 1.7B	30.5	37.0	39.2	37.5	N/A
Qwen3 0.6B	26.5	23.6	28.7	31.0	N/A
Gemma 3 1B	29.6	23.6	20.8	N/A	N/A
EXAONE 3.5 2.4B	27.2	39.3	27.8	N/A	N/A
EXAONE 4.0 1.2B	21.1	30.9	39.2	34.5	N/A

Summ

Model	8K	16K	32K	64K	128K
Mistral-Small-2506	30.0	35.7	41.5	32.7	26.7
Qwen3 235B	28.6	33.9	39.3	44.6	49.0
Qwen3 32B	26.2	32.5	37.1	40.9	43.9
Gemma 3 27B	30.7	33.1	35.1	36.5	35.9
LlaMA-4-Maverick	25.6	0.0	0.0	0.0	0.0
EXAONE 4.0 32B	22.3	24.5	27.4	26.7	27.3
SmolLM 3B	15.9	17.9	21.4	23.1	N/A
Qwen3 1.7B	15.1	17.3	20.0	22.3	N/A
Qwen3 0.6B	10.6	12.0	15.6	17.2	N/A
Gemma 3 1B	8.3	9.9	5.9	N/A	N/A
EXAONE 3.5 2.4B	17.8	19.6	19.2	N/A	N/A
EXAONE 4.0 1.2B	13.7	15.0	17.0	17.5	N/A

Figure 4: Performance of various models across six HELMET task categories, *Recall*, *RAG*, *Passage Re-ranking*, *ICL*, *LongQA*, and *Summarization*, at different context lengths (8K to 128K tokens). Darker cells indicate higher accuracy. Missing entries (`N/A`) denote models that do not support the corresponding input length or task.

D.2 RULER

We evaluate our model's long-context capabilities using the RULER benchmark [22], a synthetic evaluation suite designed to assess various aspects of long-context understanding beyond simple retrieval. RULER consists of diverse task categories, including retrieval, multi-hop tracing, aggregation, and question answering, and supports flexible configurations for context length and task complexity. The performance of our models across different sequence lengths is summarized in Table 10.

Model	4K	8K	16K	32K	64K	128K
MID-SIZE						
Mistral-Small-2506	97.18	97.15	96.66	94.57	88.53	71.84
Qwen3 235B	97.70	97.20	96.40	95.10	93.30	90.60
Qwen3 32B	98.40	96.00	96.20	94.40	91.80	85.60
Geamma3 27B	95.5	95.13	93.88	91.1	80.59	66.00
LlaMA-4-Maverick	97.10	96.85	12.96	4.90	4.35	2.85
EXAONE 4.0 32B	96.26	94.85	93.93	93.64	91.73	88.18
SMALL-SIZE						
SmolLM 3B	92.30	85.01	81.76	77.85	66.27	-
Qwen3 1.7B	89.70	86.58	80.23	75.17	$65.94^{\dagger}$	-
Qwen3 0.6B	80.74	73.64	67.17	60.82	$55.09^{\dagger}$	-
Gemma 3 1B	58.93	46.98	41.09	28.75	N/A	-
EXAONE 3.5 2.4B	88.91	87.79	87.27	77.73	N/A	-
EXAONE 4.0 1.2B	87.02	86.71	88.83	81.07	77.43	-

Table 10: Accuracy scores of MID-SIZE and SMALL-SIZE models on the RULER benchmark across varying context lengths (4K to 128K tokens). (`N/A`) indicates that models not supporting specific input lengths are omitted from evaluation. (†) denotes that models are extended to 64K context length using YaRN.

D.3 LongBench

LongBench [5] has been suggested as a bilingual benchmark to assess long context comprehension in English and Chinese. We focus on the English subsets, specifically *Single-doc QA*, *Multi-doc QA*, *Summarization*, and *Few-shot Learning*.

The *Single-doc QA* task covers datasets such as NarrativeQA [27], Qasper [7], and MultiFieldQA-EN [5]. For the *Multi-doc QA* task, we employ benchmarks including HotpotQA [64], 2WikiMultihopQA [20], and MuSiQue [57]. The *Summarization* task utilizes datasets like GovReport [23], QMSum [68], and MultiNews [11]. For the *Few-shot Learning* task, we draw from TREC [33] and TriviaQA [25]. All evaluations follow the official protocols and metrics defined in LongBench.

Comprehensive results for each task are shown in Table 11.

Context Len.	Model	Total Avg.	Single-doc QA	Multi-doc QA	Summarization	Few-shot Learning
MID-SIZE						
128K	Mistral-Small-2506	51.48	43.73	52.51	28.82	80.87
	Qwen3 235B	45.28	41.45	46.96	25.56	67.13
	Qwen3 32B	44.24	41.27	47.97	25.73	62.01
	Gemma3 27B	51.54	42.65	54.81	24.45	84.26
	LlaMA-4-Maverick	34.71	32.72	24.68	23.84	57.58
	EXAONE 4.0 32B	48.12	39.40	48.46	27.34	77.28
SMALL-SIZE						
32K	SmolLM 3B	39.85	33.38	18.26	27.94	79.83
	Qwen3 1.7B	41.82	33.61	31.87	26.16	75.63
	Qwen3 0.6B	32.72	22.75	20.29	23.11	64.73
	Gemma 3 1B	34.91	24.85	24.09	21.41	69.29
	EXAONE 3.5 2.4B	42.74	35.03	43.11	20.05	72.75
	EXAONE 4.0 1.2B	36.75	30.93	34.98	25.14	55.96
64K	SmolLM 3B	39.93	33.53	18.27	28.11	79.83
	Qwen3 1.7B†	41.92	32.01	32.53	25.95	77.19
	Qwen3 0.6B†	32.44	22.38	21.40	23.15	62.84
	Gemma 3 1B	N/A	N/A	N/A	N/A	N/A
	EXAONE 3.5 2.4B	N/A	N/A	N/A	N/A	N/A
	EXAONE 4.0 1.2B	36.93	31.02	35.09	25.28	56.33

Table 11: Task-wise performance of MID-SIZE and SMALL-SIZE models on the LongBench benchmark across four task categories: *Single-doc QA*, *Multi-doc QA*, *Summarization*, and *Few-shot Learning*. Each score represents the average accuracy over the English subset of LongBench at specified context lengths. (`N/A`) indicates that models not supporting specific input lengths are omitted from evaluation. (†) denotes that models are extended to 64K context length using YaRN.

D.4 Ko-LongBench

Ko-LongBench is an in-house benchmark developed to evaluate long-context understanding in Korean. It consists of multiple tasks, including *Document QA*, *Story Understanding*, *Dialogue History Understanding*, *In-Context Learning*, *Structured QA*, and *RAG*, allowing for a comprehensive assessment of LLMs' long-context capabilities in real-world scenarios. A detailed overview of the dataset is provided in Table 12, and representative prompt examples for each task are shown in Figures 5 and 6. Table 13 summarizes the average performance of SMALL-SIZE models on Ko-LongBench, reporting scores for both 32K and 64K context lengths.

Category	**Subtask**	**# Samples**	**Description**
SingledocQA / MultidocQA	Medical	300	Single- and Multi-Document Question Answering in the Medical Domain
	Legal	300	Single- and Multi-Document Question Answering in the Legal Domain
	Finance	300	Single- and Multi-Document Question Answering in the Finance Domain
	Patent	300	Single- and Multi-Document Question Answering in the Patent Domain
Story Understanding	Ordering	66	Evaluation of the Ability to Sequence the Given Story
	Mixeing	150	Evaluation of the Ability to Infer the number of Mixed Stories
Long-dialogue History Understanding	Wrong chatbot	150	Inferring Inconsistencies with Given Information in Multi-turn Dialogues
	Wrong inference	150	Inferring Information that cannot be deduced from Multi-turn Dialogues
	Topic classify	150	Evaluating the Ability to classify topics in Multi-turn Dialogues
Long In-context Learning	Manual QA	150	Evaluation of Information Extraction Ability based on Product Manuals
	Many-Shot	150	Evaluation of Information Extraction ability within a Few-shot Context
Long Structued QA	Table QA	300	Table-Based Question Answering : Evaluation of Table Interpretation Skills
RAG	Manual QA	150	Single-Document Question on Retrieved Document Context
	MultiQA	150	Multi-Document Question based on Retrieved Document Context
Total		**2766**	

Table 12: Descriptions of Ko-LongBench.

Ko-LongBench Example (Long-dialogue History Understanding)

다음 문제에 대해 정답을 고르세요. 당신의 최종 정답은 ABCD 중 하나이고, 정답: 뒤에 와야 합니다. 정답을 고르기 전에 차근차근 생각하고 추론하세요.

[Dialogue 0]
안녕하세요. 50대 남성입니다. 반갑습니다. 반갑습니다 저는 20대 여성입니다 그러시군요 혹시 거주하는 곳이 어디인가요? 저는 경상도에 거주하고 있어요 선생님은요? 저는 경기도에 거주하고 있어요 혹시 직업이 있으신가요? 저는 아직까지는 학생입니다. 선생님은 있으신가요? 저는 그냥 일반 직장인이랍니다. 혹시 선생님은 시험기간에 밤을 새시나요? 자주 새요.. 미리 해야하는데 ... *<중략>*

질문 : 위 대화들이 주로 다루고 있는 메인 토픽은 무엇인가?
A) 미용과 건강>건강 B) 주거와 생활 C) 개인 및 관계>연애/결혼 D) 여가와 오락>게임

정답: {*answer*}

Figure 5: Example of Long-dialogue History Understanding (Topic classification) in Ko-LongBench.

Ko-LongBench Example (Long Structued QA)

다음 문제에 대해 정답을 고르세요. 당신의 최종 정답은 ABCD 중 하나이고, 정답: 뒤에 와야 합니다. 정답을 고르기 전에 차근차근 생각하고 추론하세요.

문서 1:인천봉화초등학교 다목적교실 조성공사 시방서(건축) <table><tbody><tr><td>바탕의 종류</td><td>도장 종류</td><td>공법</td></tr><tr><td rowspan='3'>목재면, 플라스터면, 모르타르면, 콘크리트면</td><td>1종</td><td> ... *<중략>*

질문 : 2022년 경기도청 북부청사 소방시설 점검 및 소방 안전관리 대행 용역에서 본관과 별관의 면적 합계는 전체 면적의 약 몇 퍼센트를 차지하는가?
A) 약 50% B) 약 60% C) 약 70% D) 약 80%

정답: {*answer*}

Figure 6: Example of Long Structued QA (Table QA) in Ko-LongBench.

Model	Avg. up to 32K	Avg. up to 64K
SmolLM 3B	19.3	15.7
Qwen3 1.7B	62.4	57.1†
Qwen3 0.6B	18.6	16.4†
Gemma 3 1B	6.3	N/A
EXAONE 3.5 2.4B	57.8	N/A
EXAONE 4.0 1.2B	72.0	69.8

Table 13: Average performance of SMALL-SIZE models on Ko-LongBench, a multi-task benchmark designed to evaluate long-context understanding in Korean. The left column reports the average scores across all tasks up to 32K context length, while the right column shows the average scores up to 64K. (`N/A`) indicates that models not supporting specific input lengths are omitted from evaluation. (†) denotes that models are extended to 64K context length using YaRN.

D.5 WMT24++

Figure 7 presents the prompt used for LLL-as-a-judge in WMT24++ benchmark. We use the same 0-shot prompt from the official WMT24++ paper.

WMT24++ Judge Prompt

"You are a professional judge for evaluating the quality of {src_lang} to {tgt_lang} translations suitable for use in {tgt_region}. Based on the source text, the human-written translation, and machine translation surrounded with triple backticks, your task is to assess the quality of the machine translation on a continuous scale from 0 to 100. A score of 0 means "No meaning preserved," then the scale goes through "Some meaning preserved," to "Most meaning preserved and few grammatical mistakes," up to a score of 100, which means "Perfect meaning and grammar." Your output should only include the score from 0 to 100 without any additional text.

{src_lang} text: “‘{src_text}”’
{tgt_lang} human translation: “‘{tgt_text}”’
{tgt_lang} machine translation: “‘{model_text}”’

Figure 7: The judge prompt for evaluating translation quality in WMT24++ benchmark.

References

[1] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report. `https://arxiv.org/abs/2504.21318`, 2025.

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. `https://arxiv.org/abs/2412.08905`, 2024.

[3] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. `https://arxiv.org/abs/2402.14740`, 2024.

[4] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. `https://arxiv.org/abs/2305.13245`, 2023.

[5] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. `https://arxiv.org/abs/2308.14508`, 2024.

[6] Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on Uncontaminated Math Competitions, February 2025.

[7] 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.

[8] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. `https://arxiv.org/abs/2501.12948`, 2025.

[9] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. DeepSeek-V3 Technical Report. `https://arxiv.org/abs/2412.19437`, 2025.

[10] Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. `https://arxiv.org/abs/2502.12404`, 2025.

[11] 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.

[12] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. `https://arxiv.org/abs/2503.01307`, 2025.

[13] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are We Done with MMLU? `https://arxiv.org/abs/2406.04127`, 2025.

[14] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. `https://arxiv.org/abs/2408.00118`, 2024.

[15] Gemma Team. Gemma 3 Technical Report. `https://arxiv.org/abs/2503.19786`, 2025.

[16] Needle In A Haystack - Pressure Testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main`.

[17] Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork Open Reasoner 1 Technical Report. `https://arxiv.org/abs/2505.22312`, 2025.

[18] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-IF: Benchmarking LLMs on Multi-Turn and Multilingual Instructions Following. `https://arxiv.org/abs/2410.15553`, 2024.

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. `https://arxiv.org/abs/2009.03300`, 2021.

[20] 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.

[21] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020.

[22] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? `https://arxiv.org/abs/2404.06654`, 2024.

[23] 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.

[24] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. `https://arxiv.org/abs/2403.07974`, 2024.

[25] 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.

[26] Hyunwoo Ko, Guijin Son, and Dasol Choi. Understand, Solve and Translate: Bridging the Multilingual Mathematical Reasoning Gap. https://arxiv.org/abs/2501.02448, 2025.

[27] Tomáš Kočiský, 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.

[28] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. https://arxiv.org/abs/2411.15124, 2025.

[29] LG AI Ethics Principles. https://www.lgresearch.ai/about/mission#ethics.

[30] LG AI Research. EXAONE 3.0 7.8B Instruction Tuned Language Model. https://arxiv.org/abs/2408.03541, 2024.

[31] LG AI Research. EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. https://arxiv.org/abs/2412.04862, 2024.

[32] LG AI Research. EXAONE Deep: Reasoning Enhanced Language Models. https://arxiv.org/abs/2503.12524, 2025.

[33] Xin Li and Dan Roth. Learning Question Classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[34] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A Hybrid Transformer-Mamba Language Model. https://arxiv.org/abs/2403.19887, 2024.

[35] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023.

[36] The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[37] Mathematical Association of America. American Invitational Mathematics Examination. https://maa.org, 2025.

[38] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward. https://arxiv.org/abs/2405.14734, 2024.

[39] MiniMax. MiniMax-01: Scaling Foundation Models with Lightning Attention. https://arxiv.org/abs/2501.08313, 2025.

[40] MiniMax. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention. https://arxiv.org/abs/2506.13585, 2025.

[41] Mistral-AI. Magistral. https://arxiv.org/abs/2506.10910, 2025.

[42] Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open Mixture-of-Experts Language Models. https://arxiv.org/abs/2409.02060, 2025.

[43] NVIDIA. Nemotron-H: A Family of Accurate and Efficient Hybrid Mamba-Transformer Models. https://arxiv.org/abs/2504.03624, 2025.

[44] Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models. In *Forty-second International Conference on Machine Learning*, 2025.

[45] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient Context Window Extension of Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[46] 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. `https://arxiv.org/abs/2305.18290`, 2024.

[47] 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. `https://arxiv.org/abs/2311.12022`, 2023.

[48] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. `https://arxiv.org/abs/1707.06347`, 2017.

[49] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. `https://arxiv.org/abs/2402.03300`, 2024.

[50] Noam Shazeer. GLU Variants Improve Transformer. `https://arxiv.org/abs/2002.05202`, 2020.

[51] Guijin Son, Hanwool 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. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[52] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. `https://arxiv.org/abs/2104.09864`, 2023.

[53] Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The Curse of Depth in Large Language Models. `https://arxiv.org/abs/2502.05795`, 2025.

[54] Yunhao Tang and Rémi Munos. On a few pitfalls in KL divergence gradient estimation for RL. `https://arxiv.org/abs/2506.09477`, 2025.

[55] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira,

Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. `https://arxiv.org/abs/2503.19786`, 2025.

[56] Team OLMo. 2 OLMo 2 Furious. `https://arxiv.org/abs/2501.00656`, 2025.

[57] 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.

[58] 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.

[59] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. `https://arxiv.org/abs/2406.01574`, 2024.

[60] Teng Xiao, Yige Yuan, Zhengyu Chen, Mingxiao Li, Shangsong Liang, Zhaochun Ren, and Vasant G Honavar. SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters. `https://arxiv.org/abs/2502.00883`, 2025.

[61] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On Layer Normalization in the Transformer Architecture. `https://arxiv.org/abs/2002.04745`, 2020.

[62] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. `https://arxiv.org/abs/2505.09388`, 2025.

[63] Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to Nope and Back Again: A New Hybrid Attention Strategy. `https://arxiv.org/abs/2501.18795`, 2025.

[64] 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.

[65] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. `https://arxiv.org/abs/2406.12045`, 2024.

[66] Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly. `https://arxiv.org/abs/2410.02694`, 2025.

[67] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. `https://arxiv.org/abs/2503.14476`,, 2025.

[68] 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.

[69] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. `https://arxiv.org/abs/2311.07911`, 2023.

[70] Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning. `https://arxiv.org/abs/2506.01347`, 2025.