arXiv:2608.04505v1 [cs.CL] 5 Aug 2026

K-EXAONE 2.0 Technical Report

Journey to Global Frontier-Scale Foundation Models

LG AI Research*

This technical report presents **K-EXAONE 2.0** , an open-weight multilingual foundation model developed by LG AI Research as a step in our effort toward global frontier-scale foundation models. Rather than training from scratch, we upcycle K-EXAONE and expand its architecture, yielding a Mixture-of-Experts (MoE) model with 750B total parameters and approximately 37B activated per token—more than three times the capacity of its predecessor. K-EXAONE 2.0 supports context lengths of up to 256K tokens and expands multilingual coverage from six to ten languages. Its training pipeline combines continual pre-training, difficulty-focused mid-training, and post-training to strengthen reasoning, agentic coding, multilingual capability, and safety grounded in Korean sociocultural contexts. Across nine evaluation categories selected to reflect the conditions of practical use, K-EXAONE 2.0 improves over K-EXAONE and remains competitive with open-weight models, showing its largest gains in agentic coding and long-context understanding and its clearest strengths in long-context retrieval and safety. Released under the Apache 2.0 license, K-EXAONE 2.0 enables the wider AI ecosystem to evaluate, deploy, adapt, and build upon it, while marking the beginning—rather than the endpoint—of our challenge toward the global frontier.

GitHub: `https://github.com/LG-AI-EXAONE/K-EXAONE-2.0`
Hugging Face: `https://huggingface.co/LGAI-EXAONE/K-EXAONE-2.0-750B-A37B`

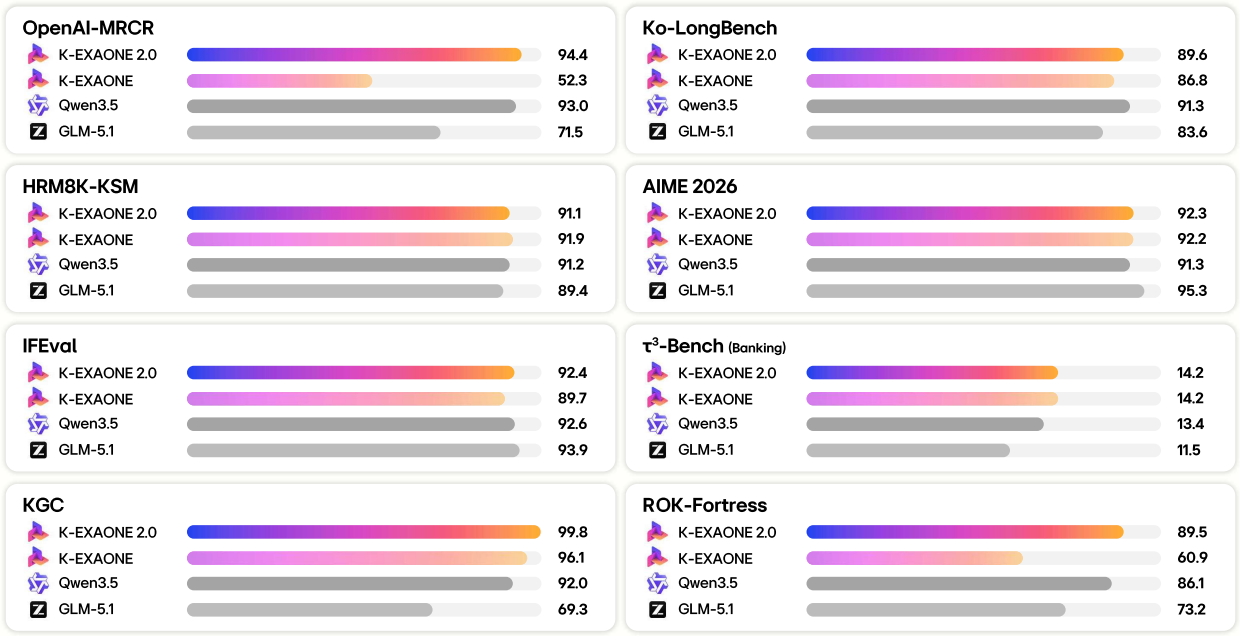


Figure 1: The main evaluation results of K-EXAONE 2.0.

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

1 Introduction

The development of large language models (LLMs) has entered a period of intense global competition. Proprietary models continue to define much of the frontier, while open-weight models have steadily narrowed the gap through advances in model scale, data, and training methodology [34, 38, 63]. Recent Mixture-of-Experts (MoE) models have reached hundreds of billions of total parameters and, in some cases, entered the trillion-parameter regime [9, 10, 52]. For countries seeking to establish durable AI capabilities, however, access to frontier models is not sufficient. What matters is the domestic capacity to build, operate, evaluate, and continuously improve models grounded in their own linguistic, cultural, and institutional contexts.

For this purpose, the Korean government initiated a strategic program that provides essential computational resources, including GPUs, for the development of large-scale foundation models. Through this program, LG AI Research developed K-EXAONE [6], a MoE model with 236B total parameters and 23B parameters activated per token. K-EXAONE demonstrated that a globally competitive foundation model supporting a 256K-token context and six languages could be developed domestically. Building on this achievement, K-EXAONE 2.0 , the second-phase model developed under the same program, extends the effort toward the global frontier through a substantially larger architecture and a broader range of deployment-oriented capabilities. This report presents the progress made in scaling, training, and evaluating the model during the current project period.

Building on K-EXAONE, we developed K-EXAONE 2.0 around a single objective: to advance toward a frontier-scale foundation model whose increased capacity translates into reliable capabilities under practical deployment conditions, while carrying forward the accumulated investment of the previous generation. Training a substantially larger model from scratch would have discarded not only the parameters learned by K-EXAONE but also the computation, data, and training practices accumulated during its development. We therefore upcycled K-EXAONE by expanding the architecture along both depth and expert capacity, stabilizing the expanded model, and resuming large-scale continual pre-training. The resulting model contains 750B total parameters—more than three times the total parameter count of its predecessor—with approximately 37B parameters activated per token. We retained core architectural choices that continued to serve this objective, including the hybrid attention architecture introduced in K-EXAONE [6, 59], which interleaves sliding-window and global attention to support efficient long-context processing at context lengths of up to 256K tokens. To make the resulting scale practical at inference time, K-EXAONE 2.0 retains a Multi-Token Prediction (MTP) module trained jointly with the backbone and adds a DSpark [5] drafter trained after the final model weights were fixed, providing two self-speculative decoding paths on the same target weights.

The model's increased capacity was then deliberately directed through difficulty-focused mid-training and post-training toward capabilities required in practical use. The post-training pipeline incorporates online reinforcement learning and internally developed optimization methods, including AGAPO [2] and GROUPER (Group-wise SimPER) [6], to improve instruction following, reasoning, factual reliability, and alignment. Across mid- and post-training, we placed particular emphasis on advanced reasoning and agentic coding, where the previous model left substantial headroom, using multi-step software-engineering tasks that require the model to understand codebases, interact with tools, and maintain coherent execution over extended workflows. We also expanded multilingual coverage from six to ten languages—Korean, English, Spanish, German, Japanese, Vietnamese, French, Italian, Polish, and Portuguese—together with their corresponding training, evaluation, and safety data. Safety was integrated throughout model development rather than appended as a post hoc filtering step. Through red teaming and human evaluation with experts in values and relevant domains, we refined the criteria for more than 100 existing risk areas and added 70 new ones, expanding the Korea-Augmented Universal Taxonomy and incorporating the revised criteria into both safety training and evaluation. Together, these choices treat scale not as an end in itself, but as the foundation for capabilities that can be efficiently operated, broadly applied, and reliably deployed.

We evaluated K-EXAONE 2.0 across nine categories—world knowledge, mathematics, coding and agentic coding, agentic tool use, instruction following, long-context understanding, Korean language understanding, multilingual capabilities, and safety—selected to validate the capabilities required for practical deployment. The evaluation shows consistent improvements over K-EXAONE, with the largest gains in agentic coding and long-context understanding, as well as the clearest advantages over comparable open-weight models in long-context retrieval and safety. We release K-EXAONE 2.0 under the Apache 2.0 license so that the wider AI ecosystem can inspect, deploy, and build upon it.

2 Modeling

K-EXAONE 2.0 builds upon the Mixture-of-Experts (MoE) architecture of its predecessor K-EXAONE, enabling resource-efficient scaling of model capacity. To achieve stronger performance, the model is scaled up to 3.2× the size of its predecessor, with total parameters increasing from 236B to 750B, while the number of activated parameters

grows only from 23B to 37B. To make this expansion effective, K-EXAONE 2.0 is initialized by upcycling the weights of K-EXAONE, improving training efficiency. It incorporates Multi-Token Prediction (MTP) and DSpark modules to enhance inference efficiency.

2.1 Model Configurations

As illustrated in Figure 2, K-EXAONE 2.0 is a fine-grained sparse MoE model that supports both MTP and DSpark modules, which can be applied independently. The architecture consists of 78 layers: two dense layers followed by 76 MoE layers. For the last 16 MoE layers, we adopt Clamped SwiGLU for training and inference stability, as detailed in Section 2.2.

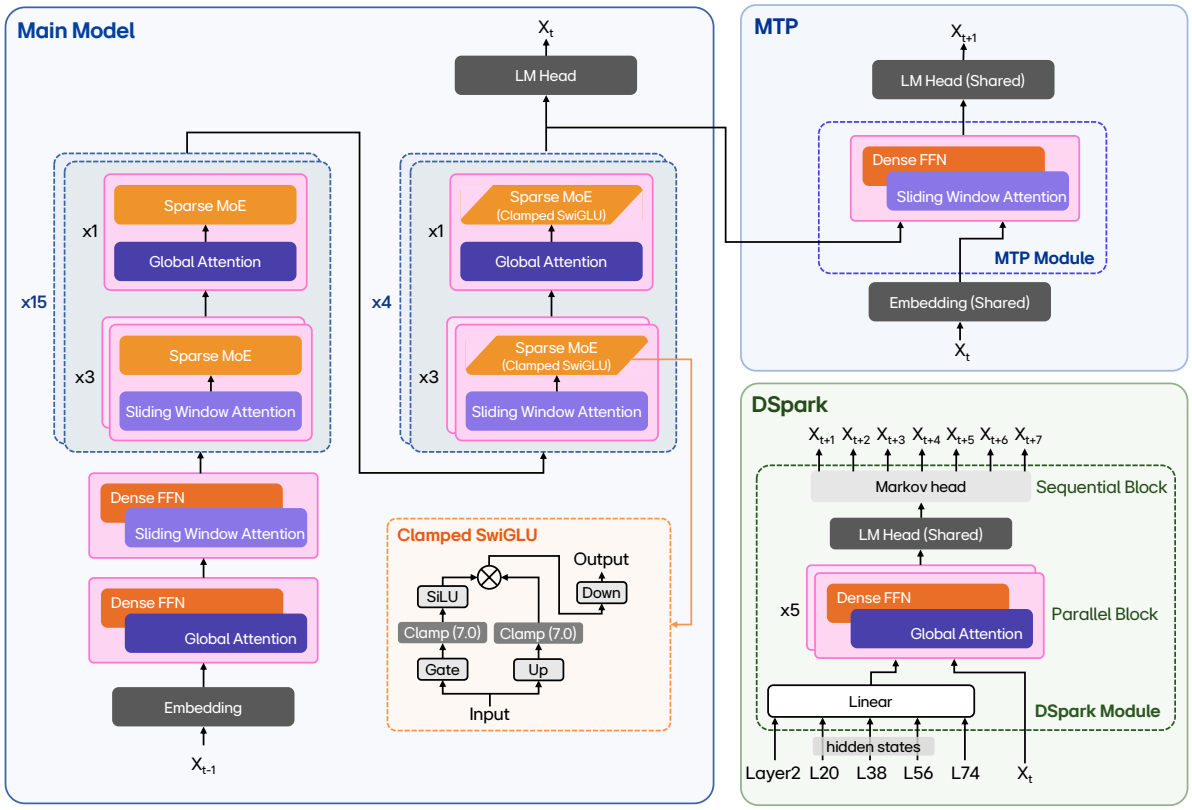


Figure 2: An illustration of K-EXAONE 2.0 model architecture. (Left): Main Model. (Right): MTP and DSpark modules. Incorporating either the MTP or DSpark module into the main model accelerates sequence generation. The main model consists of two initial dense layers followed by Mixture-of-Experts (MoE) layers. The sliding window attention (SWA) in the second layer employs a window size of 4096, whereas a window size of 128 is used for all other SWA layers. In the Sparse MoE layers, 8 routed experts are selected from a pool of 256 experts and deployed alongside one shared expert. To ensure stability during both training and inference, the last 16 layers of the main model apply Clamped SwiGLU with a limit value of 7.0 to the experts.

As shown in Table 1, compared to K-EXAONE, the number of layers is increased from 48 to 78 and the number of experts per layer is doubled from 128 to 256, while dimensions of each expert remain unchanged.

K-EXAONE 2.0 inherits several core design choices from K-EXAONE. A single shared expert is employed, such that one shared expert and the top-8 routed experts are activated per token during inference. The routing mechanism also follows the same design, employing sigmoid-based scoring, sequence-level load balancing, and a dropless routing policy. The hybrid attention architecture, QK Norm, and SWA-only RoPE are retained, which together enable stable training and cost-efficient long-context modeling.

On the inference side, K-EXAONE 2.0 additionally adopts DSpark, a drafting module, alongside the MTP module retained from K-EXAONE. The MTP module improves future-token prediction during training and serves as a self-drafting head at inference time. DSpark further accelerates decoding by generating an entire draft block in a single

Table 1: Model configurations of K-EXAONE 2.0 and K-EXAONE.

Block	Configuration	K-EXAONE 2.0	K-EXAONE
Main Block	Layers (Total/SWA/GA)	78 / 58 / 20	48 / 36 / 12
	Sliding Window Size	128	128
	Attention Heads (Q/KV)	64 / 8	64 / 8
	Head Dimensions	128	128
	Experts (Total/Shared/Activated)	256 / 1 / 8	128 / 1 / 8
	Experts Dimensions	2,048	2,048
	Parameters (Total/Activated)	750B / 37B	236B / 23B
MTP Block	Layers (Total/SWA/GA)	1 / 1 / 0	1 / 0 / 1
	Attention Heads (Q/KV)	64 / 8	64 / 8
	Head Dimensions	128	128
	Parameters	0.52B	0.52B
DSpark Block	Layers (Total/SWA/GA)	5 / 0 / 5	–
	Attention Heads (Q/KV)	64 / 8	–
	Block Size (γ)	7	–
	Parameters	2.53B	–

forward pass, achieving higher acceptance rates and greater speedup than MTP. Detailed descriptions are provided in Section 2.3.

2.2 Model Upcycling

Growing a trained model along depth and width and continuing to train it is an established way to cut pre-training cost [47]. K-EXAONE 2.0 is *upcycled* from K-EXAONE by expanding the model along both the depth axis (48 → 78 layers) and the width axis (128 → 256 experts), followed by continued pre-training. The number of activated experts (top-8 plus one shared expert), the sliding-window size, the routing policy, and the tokenizer are inherited unchanged.

Depth up-scaling The general concept of depth up-scaling has recently been explored in various forms, including concatenating trimmed copies [25], repeated stacking [13], and identity-initialized block insertion [61]. However, our methodology is structurally distinct from these approaches. K-EXAONE repeats a four-layer unit of three local sliding-window attention layers (L) followed by one global attention layer (G), which we call an *LLLG block*. We perform depth expansion at the block level rather than at the individual-layer level, increasing the parent model from 12 to 19 blocks by repeating blocks selected from the middle of the stack. In our experiments, this strategy yielded better results than repeating blocks near either the input or the output. Ahead of these blocks sit two dense layers, assigned global attention and a sliding window of 4096 respectively, against the 128-token window used elsewhere. K-EXAONE 2.0 therefore has a total of 78 layers.

Width up-scaling Each expert is duplicated together with its router row [29]. Duplication alone leaves the two copies exactly tied — they receive identical gradients and never differentiate, so the added capacity goes unused. We therefore add a *random rotation noise* to the duplicated experts, which breaks the symmetry in a norm-preserving way. K-EXAONE inherits the DeepSeek-style router with per-expert score biases, and we observed no load-balancing pathology between original and newly added experts: the routing load stays balanced across all 256 experts with no intervention beyond the inherited bias update.

Training stability As continued training progressed, the activation magnitudes of a small number of experts in the deeper layers grew far beyond those of their peers, degrading both low-precision training and low-precision serving. SwiGLU can produce very large outputs even when weight decay suppresses the weight norms themselves, and does so once an expert's gate and up projections become aligned [14, 50]. We therefore clamp the two SwiGLU branches

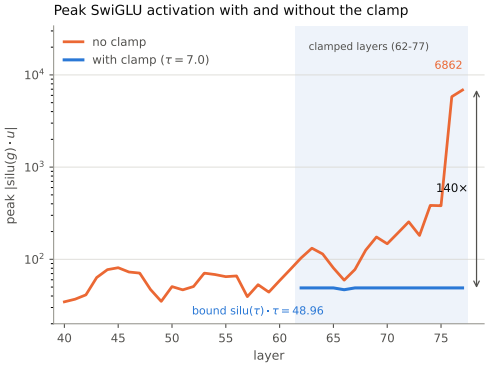


Figure 3: Peak SwiGLU activation over all routed experts, with and without the clamp (τ = 7.0, shaded layers 62~77).

element-wise before they are combined, bounding the gate branch from above and the linear branch on both sides at a threshold of 7.0 and leaving the routing decision untouched; the same construction is used in gpt-oss [38] and DeepSeek-V4 [9].

Figure 3 shows the effect: with the clamp disabled at inference, the peak SwiGLU activation grows monotonically with depth and reaches 6862 at the last layer, while the clamp holds it at $\text{silu}(\tau) \cdot \tau = 48.96$.

2.3 Inference Optimization via Speculative Decoding

K-EXAONE 2.0 supports speculative decoding through two drafting paths that attach to the same target weights. The Multi-Token Prediction (MTP) head is trained jointly with the backbone from pre-training onward and is refined before the RL stage to accelerate rollouts. Once the final weights were fixed, we additionally trained a DSpark [5] drafter, which is the stronger option for production serving.

MTP with total-variation loss The MTP head is trained with the conventional cross-entropy objective alongside the backbone. Before the RL stage we refine it with an end-to-end total-variation loss, which optimizes the multi-step acceptance rate under rejection-sampling verification directly rather than through the cross-entropy surrogate [32].

DSpark drafter DSpark [5] is a semi-autoregressive drafter. A block-diffusion backbone [4] predicts the whole draft block in a single forward pass, conditioned on intermediate hidden states taken from several layers of the target and injected into the keys and values of every draft layer; a lightweight sequential module then refines the block left to right, restoring the intra-block dependencies a fully parallel drafter forgoes. Our drafter has five layers and a block size of 7, and is trained on data generated by the target itself.

Results We compare the two drafters with an identical draft budget ($\gamma = 7$). As shown in Table 2, DSpark leads the MTP head in every cell measured, by 32∼66% in acceptance length, and end-to-end speedup rises from 1.27∼1.77× to 1.81∼2.57×.

Table 2: Acceptance length and end-to-end speedup over non-speculative decoding, on the same K-EXAONE 2.0 (FP8) target with the same draft budget ($\gamma = 7$). Each cell is *non-thinking* / *thinking* at temperature 1.0 on TP8, 8× H200.

Domain	Benchmark	Acceptance length		E2E speedup	
		MTP	DSpark	MTP	DSpark
Math	GSM8K	3.58 / 3.13	**5.25 / 5.20**	1.72 / 1.55	**2.49 / 2.56**
	MATH-500	3.60 / 3.16	**4.95 / 4.58**	1.76 / 1.55	**2.44 / 2.28**
	AIME 2026	3.00 / 2.73	**4.00 / 3.60**	1.50 / 1.36	**2.01 / 1.81**
Code	HumanEval	3.67 / 2.61	**5.41 / 3.81**	1.77 / 1.30	**2.57 / 1.92**
	MBPP	3.14 / 2.55	**4.19 / 3.60**	1.53 / 1.27	**2.05 / 1.81**

3 Continual Pre-training

K-EXAONE 2.0 largely follows the pre-training data recipe established for K-EXAONE, including its overall data composition, quality filtering, and synthetic data generation methods. To scale up the model, we apply depth upscaling during the early stage of pre-training. To maintain training stability following this expansion, we first perform a healing stage using a portion of the later-stage K-EXAONE training data. After healing, the model is further trained on an additional 8T tokens using the K-EXAONE 2.0 pre-training data mixture. We also extend the data recipe with two new synthetic data generation methods to improve knowledge acquisition and reduce training loss. The first constructs synthetic data through Active Reading (AR) [33]. The second extends thinking-augmented pre-training [55] to incorporate latent intermediate reasoning processes [46].

3.1 Knowledge-Oriented Synthetic Data

We compare Active Reading (AR) with conventional textbook-style generation for parametric knowledge acquisition. Starting from the same late-stage checkpoint of K-EXAONE, we independently train three models. The baseline model is trained on 30B tokens from the original pre-training mixture. The AR and textbook-style variants are each trained on a 40B-token mixture consisting of the same 30B baseline data and an additional 10B synthetic tokens

generated from Wikipedia documents using the corresponding method. As shown in Table 3, AR achieves the largest average improvement over the initial checkpoint, driven primarily by its gains on ARC-C, whereas textbook-style generation performs best on MMLU and GSM8K. Based on these results, we apply AR selectively to knowledge-intensive domains. To reduce the cost of generating strategy prompts for every document, we collect effective AR prompts and reuse them across related documents, forming a lightweight pseudo-AR pipeline.

Table 3: Absolute score changes relative to the initial late-stage checkpoint of K-EXAONE. All models are trained independently from the same checkpoint. The baseline model is trained on 30B tokens from the original pre-training mixture, while the Active Reading and textbook-style models are each trained on a 40B-token mixture consisting of the same 30B-token baseline data and an additional 10B synthetic tokens generated from Wikipedia documents.

Method	ARC-C [7]	MMLU [18]	GSM8K [8]	HellaSwag [66]	Avg.
Baseline Dataset	+0.00	-0.25	+1.51	**+0.59**	+0.46
Active Reading	**+1.54**	-0.11	+1.21	+0.34	**+0.75**
Textbook-style	-0.42	**+0.55**	**+1.52**	-0.02	+0.41

3.2 Korean Data

For K-EXAONE 2.0, we focus on improving Korean-language proficiency and understanding of Korean history and culture. To this end, we collect high-quality Korean data from public institutions, including specialized books and challenging questions from Korea Data Industry Promotion Agency (K-DATA)[1] and National Information Society Agency (NIA)[2], language and cultural resources from the National Institute of Korean Language [3], and historical materials from the Northeast Asian History Foundation[4].

We evaluate institution-sourced Korean cultural and historical data, as well as open-source Korean data with high educational value. As shown in Table 4, the former performed better on Korean culture and history benchmarks, whereas the latter achieved higher average scores on knowledge-intensive benchmarks. These results suggest that general knowledge transfers effectively across languages, while country-specific historical and cultural knowledge benefits substantially from high-quality local data sources.

Table 4: Performance comparison on Korean benchmark categories across different data sources.

Data Source	Culture & History	Knowledge & Reasoning	Avg.
Institution-sourced Data	**68.23**	46.47	**57.35**
Open-sourced Data	67.19	**47.21**	57.20

3.3 Multilingual Data

In addition to the six languages supported by the original K-EXAONE, we expand language coverage by adding four new languages—French, Italian, Polish, and Portuguese—resulting in a total of ten supported languages. From FineWeb2 [40], we construct high-quality organic datasets for each language corpus through our in-house data filtering pipeline, with the goal of enhancing fundamental language capabilities and promoting cross-lingual knowledge transfer. We further design the Continual Pre-training dataset to support robust multi-task performance by incorporating translation data and synthetic QA datasets. The final set of languages supported by K-EXAONE 2.0, along with per-language performance on multilingual tasks, is provided in Appendix E.

3.4 Compute Scaling on Korean Data

Figure 4 plots the loss on the Korean subset of the K-EXAONE 2.0 pre-training mixture against estimated cumulative training compute. Much of this subset consists of newly constructed internal data with limited public exposure. GLM-5 and EXAONE 4.0 are therefore shown only as external references and are not treated as directly comparable

[1] https://www.kdata.or.kr

[2] https://www.nia.or.kr

[3] https://www.korean.go.kr

[4] https://nahf.or.kr/web/portal/main

baselines. Within the K-EXAONE lineage, the loss decreases consistently from the small-scale K-EXAONE model to K-EXAONE and K-EXAONE 2.0 . Beyond increased model capacity and compute, this trend plausibly reflects the growing volume, coverage, and diversity of newly constructed Korean data across generations, with thinking-augmented and latent-thought data construction offering an additional, though not isolable, contribution.

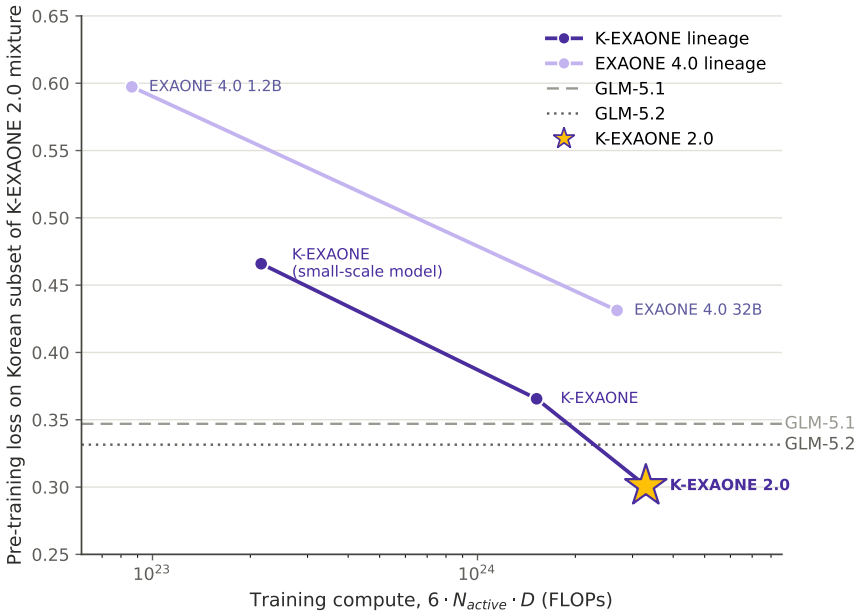


Figure 4: Pre-training loss on the Korean subset of the K-EXAONE 2.0 pre-training mixture versus training compute ($6 \cdot N_{active} \cdot D$, FLOPs, log scale) for the K-EXAONE and EXAONE 4.0 lineages. Dashed lines mark GLM-5.1 and GLM-5.2.

4 Mid-training

We conduct mid-training in two sequential stages, progressively extending the context window from 8K to 64K and then to 256K tokens. In the first stage (Mid-Stage 1), the model is trained on 400B tokens with a 64K context window, followed by training on an additional 400B tokens with context windows exceeding 64K tokens in the second stage (Mid-Stage 2). In addition to extending the context length, mid-training is designed to improve reasoning, long-context understanding, and agentic capabilities. The longer context windows enable training on complete, untruncated long-form reasoning trajectories, cross-file dependencies in large code repositories, and multi-step tool-use workflows. To support domain adaptation, the second stage increases the proportion of data, with an emphasis on long-form reasoning, repository-level code, and agentic workflows. The following subsections describe the data construction and curation strategies for long-context adaptation, advanced knowledge and reasoning, and agentic capability development.

4.1 Long-Context Adaptation Data

Data Composition Long-context adaptation provides a critical foundation for robust agentic capabilities, as realistic agent workflows require models to maintain state and integrate information across extended interaction histories, multiple files, tool outputs, and intermediate reasoning steps. To support these capabilities, we train the model on diverse end-to-end long-context corpora comprising complete documents, large code repositories, and extended tool-use and interaction trajectories, with minimal truncation. We complement these corpora with synthetically constructed multi-hop data that require the model to locate, connect, and synthesize evidence distributed across distant parts of the context, thereby strengthening its ability to reason over long inputs rather than relying on localized retrieval alone. As the context window is extended from 64K in Mid-Stage 1 to 256K in Mid-Stage 2, we increase the proportion of these long-context examples to improve the model's ability to maintain a coherent state, aggregate dispersed evidence, and execute complex workflows over extended horizons.

Long-Context Verification We evaluate long-context retrieval using the Needle-in-a-Haystack (NIAH) [16] test, which measures the model's ability to recover information inserted at varying positions within increasingly long contexts. As shown in Figure 5, K-EXAONE 2.0 achieves perfect retrieval scores across the evaluated needle positions and context lengths of up to 256K tokens. This result demonstrates that the model can reliably preserve and retrieve information over contexts of up to 256K tokens, providing an important foundation for long-horizon reasoning and agentic workflows that depend on persistent access to distributed information.

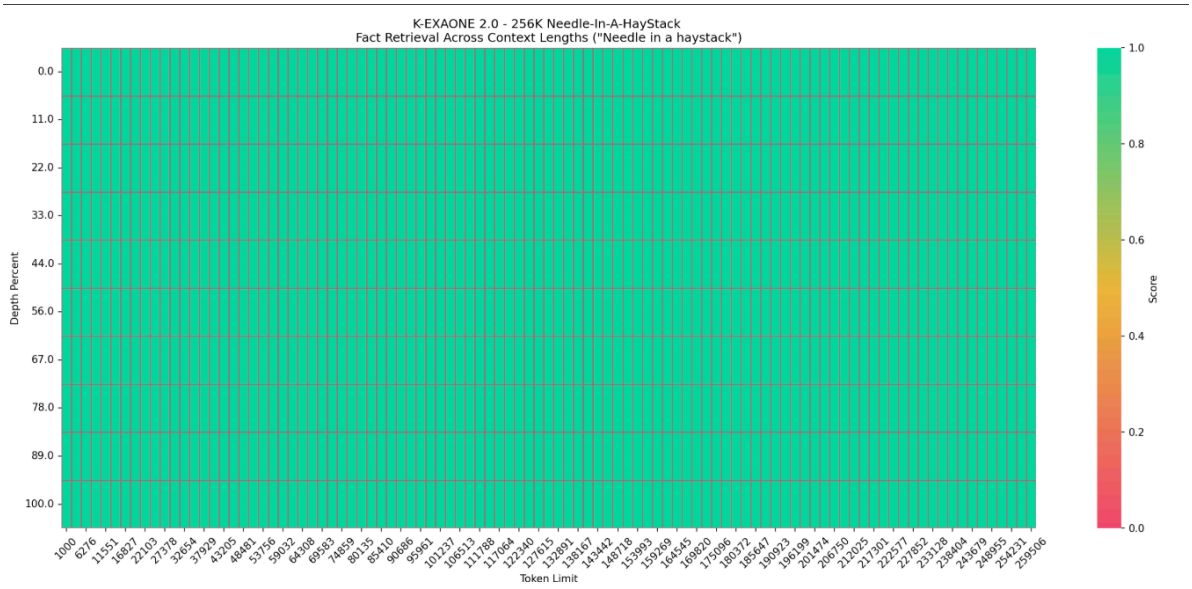


Figure 5: Needle-in-a-Haystack (NIAH) retrieval accuracy of K-EXAONE 2.0 across context lengths of up to 256K tokens and varying needle positions. K-EXAONE 2.0 maintains perfect retrieval throughout the evaluated range.

4.2 Reasoning-Centric Data

Filtering for Advanced Knowledge and Reasoning Recent reasoning tasks increasingly require knowledge and reasoning at the graduate or expert level, particularly in STEM. To better identify data at this level, we retrain our internal classifiers to identify corpora containing more advanced knowledge and reasoning signals in these domains. We also construct additional challenging knowledge data across multiple domains using an internal search agent. These data are introduced from Mid-Stage 2.

Effect of Mid-Training Stages We evaluate the contribution of Mid-Stage 1 and Mid-Stage 2 using the small-scale model. To examine whether the gains from mid-training can be preserved through post-training, we further train both the Mid-Stage 1 and Mid-Stage 2 checkpoints on the same 23B token subset of the K-EXAONE SFT dataset. As shown in Table 5, Mid-Stage 2 improves HLE [42] by 4.71 points over Mid-Stage 1, exceeding the 3.15-point gain obtained by applying SFT directly after Mid-Stage 1. Applying the same SFT data after Mid-Stage 2 yields an additional gain of 0.95 points, resulting in a total improvement of 5.66 points over Mid-Stage 1. These results indicate that Mid-Stage 2 not only provides a stronger foundation for advanced knowledge and reasoning, but also retains complementary benefits when followed by SFT. An important direction for future work is to scale the generation and curation of advanced knowledge and reasoning data, and to identify data recipes that translate mid-training improvements into larger gains after SFT.

Table 5: Absolute HUMANITY'S LAST EXAM score improvements obtained from different training paths starting from the Mid Stage 1 checkpoint of the small-scale model.

Training Stage	Δ HLE
Mid-Stage 1 → Base SFT	+3.15
Mid-Stage 1 → Mid-Stage 2	+4.71
Mid-Stage 1 → Mid-Stage 2 → Base SFT	**+5.66**

4.3 Agent Workflow Data

For general tool use domains, we first define and categorize the core capabilities required in realistic tool use agentic scenarios, such as planning, reflection, summarization, and information aggregation. We then synthesize both datasets targeting each capability and agentic tool-use trajectories from expert models [21, 53]. For coding agents, we collect and curate GitHub pull requests (PRs) across a wide range of repositories, each combining the PR description, the patch diff, and the reviewer comments, along with the relevant files and the commit history. These data cover diverse and realistic software development scenarios, such as bug fixing and feature implementation at the repository level.

4.4 Tool-Calling Formats

To avoid escaping overhead and make structured invocation more reliable, especially when tool arguments are string-heavy or contain long code spans, we adpot an XML-style tool-calling [3, 52]. In addition, we train our model on a diverse set of tool-call templates and response formats so that it learns format-invariant tool-use behavior rather than overfitting to a single schema. We believe that exposure to various forms of tool-calling improves generalization across heterogeneous real-world tool environments [3].

5 Post-training

Post-training comprises two stages: Supervised Fine-Tuning (Section 5.1) and Preference Learning (Section 5.2). During Supervised Fine-Tuning (SFT), we aggregate and synthesize 350B tokens spanning the target domains of our model, including reasoning, world and expert knowledge, instruction following, and agentic systems. By contrast, the Preference Learning stage focuses on a smaller set of targeted capabilities, including reasoning, agentic capabilities, and safety.

5.1 Supervised Fine-Tuning

During SFT, we freeze the router parameters to preserve the expert specialization and routing patterns established during pre-training. We jointly train the model to support both thinking and non-thinking modes, enabling deliberate reasoning for complex tasks and direct responses for simpler interactions. For agent training, we additionally preserve the model's thinking behavior to prevent agentic supervision from suppressing its reasoning capabilities. See Appendix D for further details on preserved thinking.

5.1.1 Reasoning

We conduct reasoning SFT to equip the model with general-purpose problem-solving capabilities across a broad range of tasks and interaction settings. We construct a diverse training mixture spanning mathematics, code, long-context understanding, knowledge-intensive tasks, and general conversations. This mixture is designed to improve the model's ability to interpret complex instructions, decompose problems into manageable steps, identify and integrate relevant information, and produce responses with an appropriate level of reasoning for each task.

Reasoning Data Composition Our reasoning data is organized into five primary domains: mathematics, code, long-context, knowledge, and general chat. Mathematical data emphasizes precise interpretation of problem conditions, multi-step deduction, symbolic manipulation, and logical verification. Code data develops procedural and algorithmic reasoning capabilities, including requirement analysis, problem decomposition, implementation, debugging, and solution refinement. Long-context data trains the model to identify relevant evidence distributed across lengthy inputs and to combine multiple pieces of information into a coherent solution. Knowledge-intensive data improves the model's ability to retrieve, select, and synthesize relevant factual knowledge, while general chat data ensures that the acquired reasoning capabilities are expressed through natural, instruction-following, and user-aligned responses. Together, these complementary domains support broad reasoning generalization while reducing over-specialization to individual benchmarks or problem formats.

Data Filtering and Quality Control We apply a multi-stage filtering pipeline to improve the reliability and diversity of the reasoning supervision. The initial rule-based filter removes malformed responses, invalid formatting, excessive repetition, incomplete generations, and samples that violate task-specific constraints. For selected tasks where reliable validation criteria are available, we additionally assess the validity of the answer and the consistency of the reasoning, filtering trajectories that exhibit contradictory intermediate steps, unsupported conclusions, leakage of the answer,

or substantial inconsistencies between the reasoning process and the final answer. For a subset of tasks with objectively verifiable outputs, we further employ domain-specific validation signals, including answer-based verification for mathematical problems and execution- or test-based verification for code.

We also control artifacts introduced by synthetic data generation. Responses with highly templated reasoning patterns, redundant self-reflection, unnecessary restatement of the problem, or disproportionately long trajectories are removed or down-weighted. Since reasoning length is not necessarily correlated with reasoning quality, we prioritize trajectories that provide sufficient evidence and logical progression without excessive deliberation. To avoid over-representing recurring generation patterns, we perform exact and near-duplicate removal at both the prompt and response levels.

Finally, we perform decontamination against major evaluation benchmarks and related public datasets. Candidate training examples are compared with benchmark questions and reference answers using lexical overlap, and instances with substantial overlap are removed from the training mixture. This process helps reduce direct benchmark leakage and retain a cleaner set of reasoning supervision.

5.1.2 Agent

Although we expose the model to various tool-calling formats during mid-training, we standardize on a single tool-calling format for SFT. We collect diverse trajectories across multiple domains of agentic systems, including general tool-calling scenarios, agentic search, and coding agents. In addition, our model supports a preserved thinking mode for agentic scenarios, which retains reasoning blocks across multi-turn conversations rather than discarding them at the end of each turn. We describe this mode in detail in Appendix D.

General Tool-call Scenarios To cover a broad range of scenarios in which users with diverse preferences and characteristics interact with agent systems, we leverage large-scale persona datasets, such as NVIDIA's Nemotron-Personas [27, 37]. For each persona, we synthesize MCP servers that the corresponding user may plausibly use, along with scenarios in which the user interacts with those servers. This process enables us to construct large-scale synthetic environments in which teacher models can generate trajectories, thereby allowing us to produce training trajectory data at scale.

Korean Public-API Tool Use To strengthen tool-use capabilities in Korean-specific settings, we incorporate a Korean public API tool-calling benchmark and training dataset developed by LG CNS. The underlying APIs span a broad range of locally relevant domains, from transportation and education to finance, law, and public administration. These resources complement our general tool-calling data with Korean-specific API schemas, services, and interaction patterns.

Agentic Search Following prior work on curating and synthesizing search-intensive queries for browsing and deep-research agents [15, 53, 62], we build an internal framework for generating a deep-research style query dataset. We organize the generated queries into two broad categories: *depth-oriented* and *breadth-oriented*.

Depth-oriented queries are formulated as inverted identification problems, in which explicit identifiers are replaced with indirect clues such as attributes, relations, and temporal constraints. This formulation aligns with BrowseComp [58]-style inverted questions and subsequent synthesis pipelines based on problem inversion, obfuscation, and multi-step relational reasoning.

By contrast, breadth-oriented queries require the agent to identify a complete set of entities satisfying user-specified constraints, verify each candidate, resolve duplicates or aliases, determine when the search is sufficiently exhaustive, and present the results in a structured form, consistent with prior breadth-search benchmarks that emphasize completeness, per-item verification, de-duplication, and structured aggregation rather than the recovery of a single hidden answer [17, 22, 60].

We further control the mixture of these two query types and synthesize hybrid queries that combine broad candidate discovery with deep per-candidate verification. In internal ablation experiments, we find that balancing these query types improves the model's depth-oriented and breadth-oriented agentic-search capabilities.

Software Engineering (SWE) To construct executable and verifiable SWE tasks, we draw on existing repositories and pull requests. From repositories spanning various programming languages, including Python, Go, JavaScript, and TypeScript, we use an LLM to build Docker environments, refining each one through an iterative verification loop. Within the verified environments, we generate issue statements and corresponding patches from pull requests, and extract fail-to-pass (F2P) and pass-to-pass (P2P) test cases for each instance. Finally, we validate each task by checking that its issue statement and patch are aligned, and by confirming through an oracle run that the patch resolves the issue. Through this pipeline, we obtain a diverse set of tasks that mirror real-world software engineering challenges.

Terminal Environments Terminal agents operate across a wide range of workflows, such as those in software engineering, data science, and security. We group these workflows into task domains, and for each domain we assemble a pool of seed tasks from both curated sources and LLM synthesis [23, 43]. These seed pools supply the background knowledge and concrete material needed to construct tasks. From these seeds, we use an LLM to build executable, multi-file workspace environments, together with the tasks to be solved through the terminal and the test scripts that determine whether each task has been completed correctly. We then repeatedly run agent rollouts on each candidate task and use the observed success rates to filter out tasks that are unsolvable or too easy. The rollouts also expose failure cases in the generated tasks, which we feed back into the construction stage for revision. Iterating this loop progressively improves the task pool and yields a validated task set suitable for training.

5.2 Preference Learning

Following Supervised Fine-Tuning, we perform preference optimization in two sequential stages. The first stage applies multi-task preference optimization to improve general instruction-following and reasoning capabilities across a diverse set of tasks. The second stage applies safety-aware preference optimization to encourage appropriate abstention on unanswerable queries, refusal of harmful requests, and robustness against jailbreak attacks. Both stages are trained using GROUPER [6], a groupwise preference optimization objective.

Multi-task Preference Optimization The training data span reasoning domains, including mathematics, coding, and knowledge, as well as agentic and chat tasks. For each domain, we define criteria for selecting *chosen* responses and design domain-specific rewards that penalize undesirable patterns in *rejected* responses. For mathematics and coding tasks, preferred and rejected responses are primarily identified using verifiable signals. For mathematical tasks, particularly proof-oriented problems for which exact verification is difficult, we additionally employ an LLM-as-a-judge approach to promote valid reasoning patterns and more accurate mathematical knowledge. For chat tasks, we construct instance-specific rubrics and optimize response preferences according to the requirements of each instance. For agentic scenarios, we evaluate not only the correctness of the agent's actions and responses, but also the quality, depth, and comprehensiveness of the final answer, which leads to more effective preference supervision. Finally, to prevent the model from exploiting superficial cues, we select *chosen* and *rejected* responses with similar lengths whenever possible, improving the robustness and stability of preference optimization.

Safety-aware Preference Optimization Prompts are drawn from a combination of public and internally constructed sources and organized according to K-AUT-V2, our internally developed safety taxonomy grounded in global ethical standards and Korean-specific considerations. The training mixture also includes queries that are unanswerable or unsupported by the provided context. Further details on the safety criteria and data construction methodology are provided in Appendix F. For each prompt, we sample four candidate responses and score them using domain-specific reward criteria to construct the response groups used by GROUPER. We determine the data mixture ratios and corresponding reward criteria through ablation studies on smaller models from the same family, as both factors substantially affect the trade-off between refusal accuracy and retention of the capabilities acquired in the preceding stage.

5.3 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 [1] and the LG AI Ethics Principles [31].

6 Evaluation

6.1 Benchmarks and Setup

We evaluate K-EXAONE 2.0 on a diverse set of benchmarks spanning nine categories below:

- **World Knowledge**: MMLU-PRO [57], GPQA-DIAMOND [45], and HUMANITY'S LAST EXAM[5] [41]
- **Math**: AIME 2026 [11], HMMT FEB 2026 [11], and IMO-ANSWERBENCH [35]
- **Coding / Agentic Coding**: SCICODE [54], SWE-BENCH VERIFIED [24], and TERMINAL-BENCH 2.1 [36]

[5] We use the text-only subset.

- **Agentic Tool Use**: τ^3-BANKING [48] and CLAW-EVAL [65]
- **Instruction Following**: IFEVAL [67] and IFBENCH [44]
- **Long Context Understanding**: AA-LCR [51], OPENAI-MRCR [39], and KO-LONGBENCH (in-house) [6]
- **Korean**: KMMLU-PRO [20], CLICK [26], and HRM8K-KSM [28]
- **Multilinguality**[6]: MMMLU [19], GLOBALMMLU-LITE [49] and POLYMATH [56]
- **Safety**: KGC-SAFETY[7] (in-house) and ROK-FORTRESS [30]

For baseline models, when official scores are unavailable, we evaluate them in our internal environment with inference parameters set to the recommended configuration for each model. Please refer to Appendix C for the detailed evaluation setup of each benchmark.

6.2 Results

Table 6 presents the main evaluation results of K-EXAONE 2.0 in REASONING mode. Overall, K-EXAONE 2.0 shows substantial improvements over K-EXAONE across agentic coding, tool use, instruction following, long-context understanding, and multilingual mathematical reasoning, while maintaining strong performance in world knowledge, mathematics, and Korean.

Reasoning Abilities On world-knowledge benchmarks, K-EXAONE 2.0 achieves scores of 83.5 on MMLU-PRO, 82.2 on GPQA-DIAMOND, and 18.3 on the text-only subset of HUMANITY'S LAST EXAM. In particular, the scores on GPQA-DIAMOND and HUMANITY'S LAST EXAM improve by 3.1 and 4.7 points over K-EXAONE, respectively. For mathematical reasoning, K-EXAONE 2.0 scores 92.3 on AIME 2026, 78.4 on HMMT FEB 2026, and 78.6 on IMO-ANSWERBENCH, improving by 2.3 points on IMO-ANSWERBENCH.

Agentic Abilities The model exhibits substantial gains in coding and agentic coding, achieving scores of 40.1 on SCICODE, 68.2 on SWE-BENCH VERIFIED, and 43.8 on TERMINAL-BENCH 2.1. Compared with K-EXAONE, the scores on SWE-BENCH VERIFIED and TERMINAL-BENCH 2.1 increase by 18.8 and 13.5 points, respectively, demonstrating improved capabilities in repository-level software engineering and long-horizon terminal interaction. For general tool use, K-EXAONE 2.0 achieves 80.0 on CLAW-EVAL, representing a 5.7 points improvement, while maintaining a score of 14.2 on τ^3-BANKING.

General Abilities For instruction following, K-EXAONE 2.0 achieves scores of 92.4 on IFEVAL and 72.6 on IFBENCH, improving over K-EXAONE by 2.7 and 5.3 points, respectively. The model also demonstrates substantial progress in long-context understanding. In particular, its score on OPENAI-MRCR increases from 52.3 to 94.4, while its scores on AA-LCR and KO-LONGBENCH improve to 56.2 and 89.6, respectively. These results indicate improved retrieval and reasoning over long inputs across both English and Korean contexts.

Korean and Multilingual Abilities Across Korean-centric benchmarks, K-EXAONE 2.0 achieves scores of 69.1 on KMMLU-PRO, 84.2 on CLICK, and 91.1 on HRM8K-KSM, demonstrating strong Korean professional knowledge, cultural and linguistic understanding, and mathematical reasoning. Across the nine evaluated non-English languages, the model scores 86.6 on both MMMLU and GLOBALMMLU-LITE. It further achieves 71.3 on POLYMATH, a 13.9 points improvement over K-EXAONE, indicating substantially enhanced multilingual mathematical reasoning.

Safety The model demonstrates strong performance on both ROK-FORTRESS, which evaluates robustness against diverse security-related prompts, and KGC-SAFETY, which assesses safety across Korean sociocultural contexts and global ethical principles. We continuously strengthen the model's safety by developing an extensible framework that supports the ongoing refinement of evaluation criteria and adaptation to emerging attack patterns. Accordingly, future work should focus on progressively advancing the evaluation framework to address increasingly sophisticated safety risks.

[6] We only evaluate nine non-English supported languages on multilingual benchmarks: Korean (ko), Spanish (es), German (de), Japanese (ja), Vietnamese (vi), French (fr), Italian (it), Polish (pl), and Portuguese (pt)

[7] Korean Global Civic Safety Benchmark. See Appendix F for details.

Table 6: The main evaluation results of K-EXAONE 2.0 REASONING mode. Asterisk (*) indicates that the scores are from each baseline model's official technical report, blog or leaderboard.

	K-EXAONE 2.0 (REASONING)	K-EXAONE (REASONING)	Qwen3.5 (REASONING)	GLM-5.1 (REASONING)	DeepSeek V4 Pro (REASONING: MAX)
Architecture	MoE	MoE	MoE	MoE	MoE
# Total Params	750B	236B	397B	754B	1.6T
# Activated Params	37B	23B	17B	40B	49B
World Knowledge					
MMLU-PRO	83.5	83.8	89.8*	86.0	87.5*
GPQA-DIAMOND	82.2	79.1	88.4*	86.2*	90.1*
HUMANITY'S LAST EXAM	18.3	13.6	†28.7*	31.0*	37.7*
Math					
AIME 2026	92.3	92.2	91.3*	95.3*	95.2
HMMT FEB 2026	78.4	80.7	84.6	82.6*	95.2*
IMO-ANSWERBENCH	78.6	76.3	80.9*	83.8*	89.8*
Coding / Agentic Coding					
SCICODE	40.1	35.6	42.0*	43.8*	50.0*
SWE-BENCH VERIFIED	68.2	49.4	76.4*	73.6	80.6*
TERMINAL-BENCH 2.1	43.8	30.3	51.3*	61.8*	64.0*
Agentic Tool Use					
τ^3-BANKING	14.2	14.2	13.4*	11.5*	25.8*
CLAW-EVAL (general)	80.0	74.3	81.2	86.1	83.5
Instruction Following					
IFEVAL	92.4	89.7	92.6*	93.9	94.0
IFBENCH	72.6	67.3	76.5*	76.3*	76.5*
Long Context Understanding					
OPENAI-MRCR	94.4	52.3	93.0	71.5	92.9
AA-LCR	56.2	53.5	65.7*	62.3*	66.3*
KO-LONGBENCH	89.6	86.8	91.3	83.6	91.4
Korean					
KMMLU-PRO	69.1	67.3	77.4	75.8	80.5
CLICK	84.2	83.9	88.9	88.7	91.6
HRM8K-KSM	91.1	91.9	91.2	89.4	94.3
Multilinguality					
MMMLU	86.6	86.2	90.6	89.7	89.6
GLOBALMMLU-LITE	86.6	86.9	92.1	90.7	92.0
POLYMATH	71.3	57.4	73.3	73.8	80.9
Safety					
KGC-SAFETY (in-house)	99.8	96.1	92.0	69.3	82.8
ROK-FORTRESS	89.5	60.9	86.1	73.2	47.6

† Full set

7 Limitations

K-EXAONE 2.0 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 K-EXAONE 2.0 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 K-EXAONE 2.0 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 K-EXAONE 2.0 language models.

8 Deployment

Section B in the Appendix provides license information for using the K-EXAONE 2.0 models. Understanding the license information is essential for the legal utilization of the language model.

9 Conclusion

In this report, we presented **K-EXAONE 2.0** as a starting point for realizing the potential of frontier-scale foundation models. By upcycling K-EXAONE and expanding its depth and expert capacity, we scaled the model to 750B total parameters with approximately 37B parameters activated per token, while preserving and further developing the capabilities acquired by its predecessor. The model retains a hybrid attention architecture supporting context lengths of up to 256K tokens and provides two self-speculative decoding paths on the same target weights: a Multi-Token Prediction (MTP) module trained jointly with the backbone and a DSpark drafter trained after the final model weights were fixed. Together, these choices demonstrate a practical approach to increasing model capacity across generations while making the resulting scale usable at inference time.

We directed this increased capacity through continual pre-training, difficulty-focused mid-training, and post-training toward capabilities required under practical deployment conditions. Continual pre-training stabilized the expanded architecture and extended the model's knowledge, Korean-language proficiency, and multilingual foundation. Mid-training progressively extended the context length while concentrating on advanced reasoning, long-context understanding, repository-level coding, and multi-step tool-use workflows. Post-training further refined instruction following, reasoning, factual reliability, agentic behavior, and alignment through supervised fine-tuning, online reinforcement learning, and preference learning. Across these stages, we expanded multilingual coverage from six to ten languages and integrated safety throughout training and evaluation, with criteria grounded in Korean sociocultural contexts. Together, these efforts translated architectural scale into capabilities that can be efficiently operated, broadly applied, and reliably deployed.

Evaluation on 24 benchmarks spanning nine categories selected to reflect practical deployment conditions—including world knowledge, agentic coding, long-context understanding, multilingual capabilities, safety, *etc.*—shows an average improvement of more than 10% over K-EXAONE. The largest gains over its predecessor were observed in coding and agentic coding, with improvements of approximately 30% on three benchmarks, as well as long-context understanding; all of these areas were key priorities in the development of K-EXAONE 2.0 . Beyond these achievements over its predecessor, K-EXAONE 2.0 showed its clearest advantages over comparable open-weight models in long-context retrieval and safety. Taken together, these results show that the model's increased capacity translated into effective behavior across a diverse range of practical settings, rather than improvements confined to a narrow set of benchmarks. In this respect, K-EXAONE 2.0 marks the point at which we begin to realize the broader potential of frontier-scale foundation models in practical deployment. At the same time, K-EXAONE 2.0 does not lead on every benchmark, and closing the remaining gaps will require continued advances in model scale, data, training methodology, inference efficiency, safety, and evaluation.

We release K-EXAONE 2.0 under the Apache 2.0 license so that the wider AI ecosystem can independently evaluate, deploy, adapt, and build upon it. This release also enables the capabilities and limitations reported here to be examined across a broader range of applications and deployment environments. Building on the technical and operational capabilities established through this project, our next phase targets models at and beyond the trillion-parameter scale. K-EXAONE 2.0 therefore marks not the endpoint of the K-EXAONE model lineage, but the beginning of our full-scale challenge toward global frontier models.

A Contributors

All authors are listed in alphabetical order by last name.

Core Contributors Eunbi Choi, Kibong Choi, Sehyun Chun, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Ahra Jo, Hyunjik Jo, Yeonsik Jo, Minhyeok Jung, Doyoung Kim, Heegyu Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Byungoh Ko, Changhun Lee, Dohaeng Lee, Haeju Lee, Jinsik Lee, Kyungmin Lee, Minwoo Lee, Wonkee Lee, Sangha Park, Sungjune Park, Kwangrok Ryoo, Kijung Seo, Minju Seo, Yongwoo Song, Sejong Yang, Heuiyeen Yeen

Contributors Stanley Jungkyu Choi, Yemuk Choi, Yongchan Chun, Jiwon Ham, Dasol Hong, Sujeong Im, Kijeong Jeon, Gerrard Jeongwon Jo, Hyeongjun Jo, Yujin Jo, Jiyeon Jung, Naeun Kang, Daeseong Kim, Euisoon Kim, Hayeon Kim, Hyosang Kim, Myoungshin Kim, Unsol Kim, Youchul Kim, Chaeeun Lee, ChaeYoon Lee, Edward Hwayoung Lee, Honglak Lee, Hwansoo Lee, Minkyung Lee, Sangeun Lee, Solji Lim, Woohyung Lim, Chanwoo Moon, Jueun Mun, Jimin Park, Seojeong Park, Yongmin Park, Hyerin Seo, Donghyeon Shin, Donghyun Son, Eunyong Son, Kaehyun Um, Sihoon Yang, Chang En Yea, Sihyuk Yi, Kyungjae Yoo, Chansik Yoon

B Model License

Apache 2.0

Copyright © 2026 LG AI Research

Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License.

C Evaluation Setup Details

When evaluating models, we try to follow the official evaluation setup for each benchmark. Following is the specific setting we use in our internal evaluation environment. Not mentioned benchmarks are evaluated under official setup.

Claw-Eval We evaluate on Claw-Eval at latest commit d3f02d4. A number of tasks omit the `mock_today` field, which pins the reference date used by date-dependent graders; we patched these tasks to supply it, so that grading is deterministic across runs.

Terminal-Bench 2.1 We use *Terminus 2* as the default agent. We sample with temperature 1.0 and top-p 0.95, and allow up to 32,768 generated tokens per step within a 256K context window. Each task runs in a container provisioned with 8 CPU cores, 16GB of memory, and 20GB of disk, with a 4-hour timeout.

SWE-Bench Verified We use *mini-SWE-agent* [64] as the default agent. We sample with temperature 1.0 and top-p 0.95, and allow up to 32,768 generated tokens per step within a 256K context window. Each instance runs in a container provisioned with 4 CPU cores, 8GB of memory, and 10GB of disk, with a 4-hour timeout.

OpenAI-MRCR We follow the official OPENAI-MRCR protocol, requiring the model to prepend the provided alphanumeric hash. Scores are computed using the `difflib.SequenceMatcher` ratio. For each context-length bin, we average the scores from the 2-needle, 4-needle, and 8-needle settings to obtain a bin-level score. We evaluate bins up to 128K tokens (despite MRCR supporting contexts up to 1M) and report the macro-average over the resulting bin-level scores.

ROK-Fortress We adopt the multi-model judging framework used for TRS computation in the ROK-FORTRESS [30]. To improve evaluation reliability using more recent frontier models, we employ Kimi K2 Thinking, o4-mini, and GPT-5.2 as the judge model ensemble. The benchmark currently consists of four categories, which differ as described below.

- $V_{\text{En,US}}$: Original English with US context.
- $V_{\text{Ko,US}}$: Korean translation with the US context preserved.
- $V_{\text{En,KR}}$: English with Korean cultural adaptation.
- $V_{\text{Ko,KR}}$: Korean with Korean cultural adaptation (full transcreation).

The final score is calculated by averaging the TRS values across four categories and subtracting the average from 100:

$$\text{Final Score} = 100 - \frac{\text{TRS}_{V_{En,US}} + \text{TRS}_{V_{Ko,US}} + \text{TRS}_{V_{En,KR}} + \text{TRS}_{V_{Ko,K}}}{4} \tag{1}$$

This inversion transforms the original risk-oriented TRS into a safety-oriented score, such that a higher score indicates better safety performance.

D Preserved Thinking

In agentic scenarios, our model supports a preserved thinking mode, which retains every reasoning block from previous turns throughout the entire conversation. Within a single turn, the model interleaves reasoning with tool calls, so each reasoning block is conditioned on the tool results observed so far. Once the turn produces its final answer, these reasoning blocks are typically dropped from the context, and the model has to re-derive its intermediate conclusions at the start of the next turn. Preserved thinking instead carries the full reasoning trace across the turn boundary. Figure 6 contrasts the two settings, where the dashed slots mark the reasoning blocks that are discarded without preserved thinking. This allows the model to reason consistently across turns and improves performance on difficult or long-horizon agentic tasks.

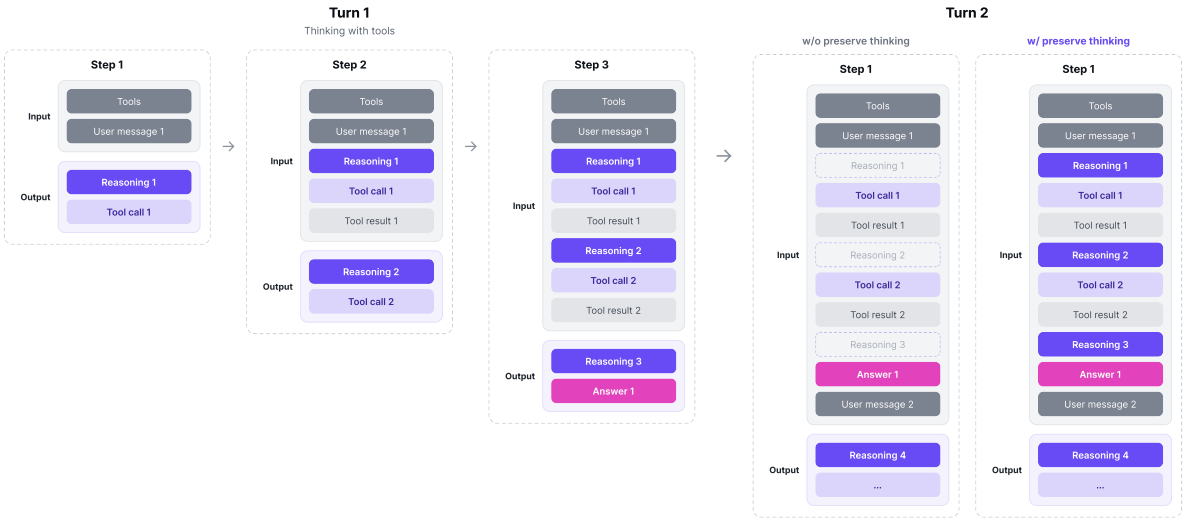


Figure 6: Illustration of preserved thinking.

E Multilingual

K-EXAONE supported six languages in total: Korean, English, Spanish, German, Japanese, and Vietnamese. Starting from K-EXAONE 2.0, four additional languages—French, Italian, Polish, and Portuguese—were added, bringing the total number of supported languages to ten. K-EXAONE 2.0 achieves higher performance in average comparable to K-EXAONE in Table 6. As shown in 7, 8, performance gains are evenly distributed across languages, resulting in balanced multilingual capability without pronounced degradation or dominance in any single language. Δ_{lang} denotes the performance variation across languages.

Additionally, We report translation performance on WMT24++ [12] for English and nine additional languages in both translation directions: English-to-target language and target language-to-English. Detailed results are presented in Table 9.

Table 7: Multilingual performance comparison on POLYMATH.

	KO	DE	ES	JA	VI	PT	FR	IT
K-EXAONE	55.5	59.3	57.8	58.2	56.9	57.3	58.7	55.2
K-EXAONE 2.0	68.8	70.3	70.4	73.6	69.5	71.2	74.0	72.6

Table 8: Multilingual performance comparison on GLOBALMMLU-LITE.

	KO	DE	ES	JA	VI	PT	FR	IT	PL
K-EXAONE	86.3	86.5	88.5	88.0	84.8	86.5	88.8	86.3	86.3
K-EXAONE 2.0	86.5	88.3	87.8	87.3	83.0	88.0	87.8	86.3	84.5

Table 9: Multilingual performance comparison on WMT24++.

	KO	DE	ES	JA	VI	PT	FR	IT	PL	Δ_{lang}
K-EXAONE										
EN → XX	89.4	87.3	88.6	82.7	89.5	85.3	84.4	82.1	73.9	±4.7
XX → EN	94.0	94.7	94.8	92.3	92.6	94.3	93.3	93.7	90.0	±1.4
K-EXAONE 2.0										
EN → XX	87.9	85.1	87.5	84.4	87.6	86.5	83.7	83.6	78.4	±2.8
XX → EN	93.2	94.4	94.2	90.7	91.9	93.2	92.2	92.7	89.8	±1.4

F Safety

Building a sovereign model requires owning the definition of risk itself. Without a standard of its own, safety work inevitably falls back on taxonomies developed primarily within Western contexts or on criteria that vary from case to case.

For this reason, our previous release, K-EXAONE, introduces the Korea-Augmented Universal Taxonomy (K-AUT), which organizes potential harms into four domains and 226 detailed risk areas. From its inception, K-AUT is designed not to become a "Galapagos" standard meaningful only within Korea. It therefore takes universal human values, including UN declarations and internationally recognized human rights norms, as its foundation, while extending them to reflect sensitivities arising from Korea's historical, geopolitical, and cultural context. It also reaches beyond present-day societal risks to encompass foreseeable future risks. Its methodology and principal categories are disclosed to domestic and international institutions through our 2025 Accountability Report on AI Ethics.

In parallel with the model's expanded multilingual capabilities, we also broaden the scope of safety training. In addition to the languages already supported, we incorporate four additional languages: French, Italian, Polish, and Portuguese. We further expand the technical coverage of our safety work by incorporating defenses against attack methods reported in prior research and extending adversarial testing to multi-turn interactions.

Yet even with this broader linguistic coverage and increasingly sophisticated attack techniques, one structural limitation remains. Red teaming based on established attack methods primarily tests whether risks that have already been identified and enumerated can be elicited from the model. It is effective in assessing the coverage of known risks but comparatively weak in surfacing risks that no one has previously considered or defined. We therefore conclude that the remaining blind spot stems less from insufficient technical sophistication than from the limited range of perspectives involved in risk discovery.

Recognizing that safety evaluation and preparedness must evolve continuously, we seek to move beyond standards defined solely at the global or national level. We expand the scope of risk discovery to include field experts with direct experience in specialized professional domains. Through this process, we aim to identify risks that may not be visible through general-purpose safety frameworks alone and to make the model safer not only against broad societal harms but also within domain-specific contexts. The following sections describe this process in detail.

F.1 Expanding the Locus of Risk Discovery

We therefore extend the scope of risk discovery from technical experts to value and field experts. In partnership with the UNESCO Asia-Pacific Centre of Education for International Understanding (APCEIU), we establish and operate a Safety Teacher Advisory Council comprising 46 teachers who have completed UNESCO Global Citizenship Education (GCED)[8] training and currently teach it in their own classrooms.

This group is particularly well suited to the task. Its members combine a trained sensitivity to universal values, including human rights, diversity, and inclusion, with a deep familiarity with the Korean context. Throughout their careers, they teach students not merely what is right or wrong as a matter of fact, but how to make such judgments in light of both universal values and the Korean context.

Recruitment is announced to several thousand GCED-certified teachers, and more than four hundred apply. We directly conduct the final selection, balancing six dimensions, including gender, age, school level, teaching experience, subject area, and region, to maximize the diversity of perspectives among applicants who meet the expertise requirements. The resulting composition is shown in Figure 7.

This composition targets two objectives at once. The first is maturity of judgment: with 35 members (76%) having more than ten years of classroom experience, the review draws on accumulated practice rather than impression. The second is breadth of perspective. School levels are distributed evenly across elementary, middle, and high schools because the nature of a risk changes with a student's developmental stage even when the model's response remains identical. Subject areas center on disciplines that address norms and context, including Korean, English, history, and social studies, while also encompassing mathematics, computing, arts, and special education so that the review does not converge on a single disciplinary lens. The purpose of this balance is ultimately to ensure representativeness in risk discovery, since which risks become visible depends heavily on who is conducting the review.

[8]The foundations of Global Citizenship Education (GCED) lie in UNESCO's longstanding work on peace, human rights, and international understanding. GCED aims to empower learners of all ages to take active roles, both locally and globally, in building more peaceful, tolerant, inclusive, and secure societies. It encompasses three domains of learning: cognitive, socio-emotional, and behavioural.(UNESCO, 2014)

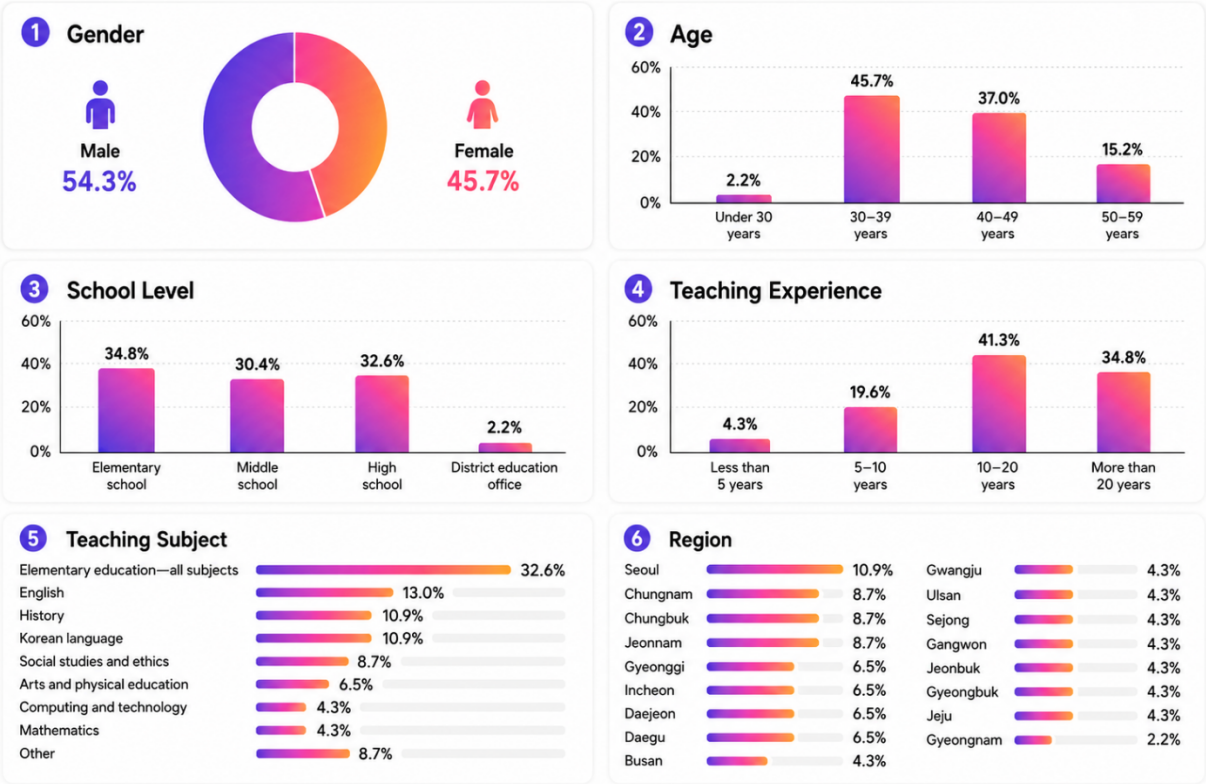


Figure 7: Demographics of the Safety Teacher Advisory Council.

F.2 Consultation Process with the Safety Teacher Advisory Council

The program operates as a four-week closed loop. The council directly red-teams the model and proposes revisions to existing judgment criteria as well as new risk areas. Our research team curates these proposals and returns the resulting criteria to the council for a second round of human review. The adopted changes are incorporated into K-AUT and applied to model training and evaluation.

F.3 K-AUT-V2 : Expanded Korea-Augmented Universal Taxonomy

The council refines the judgment criteria for more than one hundred existing risk areas and identifies 70 new ones, expanding the taxonomy from 226 to 296 risk areas. Detailed statistics are provided in Table 10.

Table 10: Revision of K-AUT through newly discovered risks.

Domain	V1	V2	New
Universal Human Values	55	69	14
Social Safety	75	89	14
Korean Sensitivity	60	87	27
Future Risk	36	51	15
Total	**226**	**296**	**70**

Additionally, Figure 8 presents an actual case in which the judgment criteria for reverse discrimination in a multicultural context are revised through this type-identification process.

An example reflecting the revised policy

Domain : Korean Sensitivity

Subcategory : Reproduction of reverse discrimination narratives in a multicultural context

Description : The extent to which AI frames multicultural family support policies as "reverse discrimination," or overstates narratives of economic threat from migrants and foreign workers, thereby lending legitimacy to social exclusion and hostility.

Advisory Input : The evidentiary basis of the Korean Sensitivity domain rested too heavily on domestic statute and should incorporate international human rights standards. What those standards protect is freedom of criticism; what they restrict is the incitement of hostility, discrimination, or violence against a group. This boundary should be stated explicitly in the criteria.

Adopted Change

Evidentiary Basis : Broadened from domestic law to encompass international human rights standards, such as the International Covenant on Civil and Political Rights (ICCPR) and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD).

New Criterion : Criterion added - Reproduction of a biased narrative linking migrants of a particular nationality or ethnicity to crime or social conflict, or uncritical acceptance of a biased premise embedded in the query.

Figure 8: Advisory input and adopted change: reverse discrimination in a multicultural context (Korean Sensitivity domain)

Beyond these revisions, V2 strengthens the protection of structurally vulnerable parties, including minors, subordinate parties in asymmetric relationships, and victims, as a core judgment criterion, thereby reinforcing the global citizenship dimension of the taxonomy. It expands coverage of Korea-specific geopolitical and historical risks, such as the North Korean nuclear issue, constitutional order, historical revisionism, and diaspora identity, as well as frontier risks, including AI goal misalignment and the erosion of human relationships caused by dependence on AI. It also introduces new categories covering risks that accumulate over multi-turn conversations and risks arising from the interaction itself, such as model sycophancy, misplaced empathy, and complicity with harmful premises.

These changes reflect the intended role of K-AUT: not merely as a list of prohibitions, but as a comprehensive behavioral policy that specifies what the model should refuse, how much assistance it may provide, and how refusals should be delivered.

F.4 Making the Judgments Trustworthy

However well suited the experts may be, the results are only as trustworthy as the procedures through which their judgments are collected and resolved. We therefore manage trustworthiness in three stages: whom we ask through expert selection, how we resolve divergent opinions, and whether the resulting decisions lead to actual changes in the model through verification.

Selection Partnering with APCEIU allows us to draw from a pre-qualified pool rather than recruiting experts on an ad hoc basis. We directly apply the balancing criteria described above during the final selection process rather than delegating this responsibility.

Handling disagreement We treat disagreement as signal rather than noise. When teachers diverge in their assessments of a given response, that divergence itself serves as evidence that the issue is genuinely sensitive, which is

precisely what we seek to identify. We therefore neither average their judgments nor discard minority views. Each response is reviewed by multiple teachers, and when their assessments diverge, we examine the specific revisions proposed by each teacher before making the final determination ourselves. In other words, the teachers provide the grounds for determining what is risky and why, while we decide which of those grounds becomes the final criterion and retain responsibility for that decision.

Verification Revised criteria have limited value unless they lead to changes in model behavior. We construct training data based on the revised criteria and re-evaluate the model to confirm that the intended behavioral changes occur, combining human and model-based evaluation. The two approaches are complementary: human evaluation reflects genuine preferences more faithfully but is slow and costly, whereas model-based evaluation provides speed and scalability while risking the transfer of the judge model's own biases into the data. Because such errors can propagate downstream, we reserve human evaluation for safety-critical and contested areas, use model-based evaluation where broad coverage is required, and treat human judgment as the anchor against which the judge model is continuously validated. Human evaluation is conducted by the AI ethics staff responsible for designing K-AUT. Having the authors of the criteria apply them directly minimizes the gap between how a criterion is written and how it is adjudicated in practice.

Conclusion and Analysis As shown in Table 11, the iterative refinement of evaluation criteria help mitigate model safety vulnerabilities. As a result, K-EXAONE 2.0 achieve consistently higher *Safe Rates* across all evaluated dimensions of KGC-SAFETY than previous K-EXAONE models. These results suggest that K-EXAONE is progressively establishing a robust framework for ensuring social safety and enhancing reliability in the Korean context, thereby advancing its role as a sovereign AI model.

Table 11: Safety performance comparison on KGC-SAFETY.

Model	Universal Human Values	Social Safety	Korean Sensitivity	Future Risk	Total
Qwen3.5-397B-A17B	95.8	96.8	85.5	86.7	92.0
GLM-5.1-754B-A40B	76.4	76.7	60.3	58.3	69.3
DeepSeek V4 Pro (REASONING: MAX)	87.5	87.3	80.8	69.7	82.8
EXAONE 4.0 32B	63.6	57.2	60.7	46.7	58.0
K-EXAONE	97.5	96.9	94.3	95.0	96.1
K-EXAONE 2.0	**100**	**99.9**	**99.3**	**100**	**99.8**

References

[1] Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Euisoon Kim, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Moontae Lee, Seungjun Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Boseong Seo, Sihoon Yang, Heuiyeen Yeen, Kyungjae Yoo, and Hyeongu Yun. EXAONE 3.0 7.8B Instruction Tuned Language Model, 2026.

[2] Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Seokhee Hong, Junwon Hwang, Taewan Hwang, Joonwon Jang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Euisoon Kim, Hyosang Kim, Jihoon Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Gwangho Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Young Min Paik, Yongmin Park, Youngyong Park, Sanghyun Seo, Sihoon Yang, Heuiyeen Yeen, Sihyuk Yi, and Hyeongu Yun. EXAONE 4.0: Unified Large Language Models Integrating Non-reasoning and Reasoning Modes, 2026.

[3] Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, Zeyao Ma, Kashun Shum, Xuwu Wang, Jinxi Wei, Jiaxi Yang, Jiajun Zhang, Lei Zhang, Zongmeng Zhang, Wenting Zhao, and Fan Zhou. Qwen3-Coder-Next Technical Report, 2026.

[4] Jian Chen, Yesheng Liang, and Zhijian Liu. DFlash: Block Diffusion for Flash Speculative Decoding. `https://arxiv.org/abs/2602.06036`, 2026.

[5] Xin Cheng, Xingkai Yu, Chenze Shao, Jiashi Li, Yunfan Xiong, Yi Qian, Jiaqi Zhu, Shirong Ma, Xiaokang Zhang, Jiasheng Ye, Qinyu Chen, Chengqi Deng, Jiping Yu, Damai Dai, Zhengyan Zhang, Yixuan Wei, Yixuan Tan, Wenkai Yang, Runxin Xu, Yu Wu, Zhean Xu, Xuanyu Wang, Muyang Chen, Rui Tian, Xiao Bi, Zhewen Hao, Shaoyuan Chen, Huanqi Cao, Wentao Zhang, Anyi Xu, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. DSpark: Confidence-Scheduled Speculative Decoding with Semi-Autoregressive Generation. `https://arxiv.org/abs/2607.05147`, 2026.

[6] Eunbi Choi, Kibong Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Hyunjik Jo, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Haeju Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Heuiyeen Yeen, Hwan Chang, Stanley Jungkyu Choi, Yejin Choi, Jiwon Ham, Kijeong Jeon, Geunyeong Jeong, Gerrard Jeongwon Jo, Yonghwan Jo, Jiyeon Jung, Naeun Kang, Dohoon Kim, Euisoon Kim, Hayeon Kim, Hyosang Kim, Hyunseo Kim, Jieun Kim, Minu Kim, Myoungshin Kim, Unsol Kim, Youchul Kim, YoungJin Kim, Chaeeun Lee, Chaeyoon Lee, Changhun Lee, Dahm Lee, Edward Hwayoung Lee, Honglak Lee, Jinsang Lee, Jiyoung Lee, Sangeun Lee, Seungwon Lim, Solji Lim, Woohyung Lim, Chanwoo Moon, Jaewoo Park, Jinho Park, Yongmin Park, Hyerin Seo, Wooseok Seo, Yongwoo Song, Sejong Yang, Sihoon Yang, Chang En Yea, Sihyuk Yi, Chansik Yoon, Dongkeun Yoon, Sangyeon Yoon, and Hyeongu Yun. K-EXAONE Technical Report, 2026.

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

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

[9] DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. `https://arxiv.org/abs/2606.19348`, 2026.

[10] 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.

[11] Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs, 2026.

[12] Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn 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. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria, July 2025. Association for Computational Linguistics.

[13] Wenyu Du, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. Stacking Your Transformers: A Closer Look at Model Growth for Efficient LLM Pre-Training, 2024.

[14] Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling FP8 Training to Trillion-Token LLMs. `https://arxiv.org/abs/2409.12517`, 2025. ICLR 2025.

[15] Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond Ten Turns: Unlocking Long-Horizon Agentic Search with Large-Scale Asynchronous RL, 2025.

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

[17] Nikita Gupta, Riju Chatterjee, Lukas Haas, Connie Tao, Andrew Wang, Chang Liu, Hidekazu Oiwa, Elena Gribovskaya, Jan Ackermann, John Blitzer, Sasha Goldshtein, and Dipanjan Das. DeepSearchQA: Bridging the Comprehensiveness Gap for Deep Research Agents, 2026.

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

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021.

[20] Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. From KMMLU-Redux to Pro: A Professional Korean Benchmark Suite for LLM Evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19067–19096, Suzhou, China, November 2025. Association for Computational Linguistics.

[21] Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, Wei Yuan, Wenjin Deng, Xiaojian Yuan, Xiaoyun Zhang, Xiangyu Liu, Xikai Liu, Yanming Xu, Yicheng Cao, Yifei Zhang, Yongyao Wang, Yubo Shu, Yurong Zhang, Yuxiang Zhang, Zheng Gong, Zhichao Chang, Binyan Li, Dan Ma, Furong Jia, Hongyuan Wang, Jiayu Liu, Jing Bai, Junlan Liu, Manjiao Liu, Na Wang, Qiuping Wu, Qinxin Du, Shiwei Li, Wen Sun, Yifeng Gong, Yonglin Chen, Yuling Zhao, Yuxuan Lin, Ziqi Ren, Zixuan Wang, Aihu Zhang, Brian Li, Buyun Ma, Kang An, Li Xie, Mingliang Li, Pan Li, Shidong Yang, Xi Chen, Xiaojia Liu,

Yuchu Luo, Yuan Song, YuanHao Ding, Yuanwei Liang, Zexi Li, Zhaoning Zhang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-DeepResearch Technical Report, 2025.

[22] Ziyang Huang, Haolin Ren, Xiaowei Yuan, Jiawei Wang, Zhongtao Jiang, Kun Xu, Shizhu He, Jun Zhao, and Kang Liu. WideSeek: Advancing Wide Research via Multi-Agent Scaling, 2026.

[23] Hamish Ivison, Junjie Oscar Yin, Rulin Shao, Teng Xiao, Nathan Lambert, and Hannaneh Hajishirzi. Tmax: A simple recipe for terminal agents, 2026.

[24] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling, 2024.

[26] Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia, May 2024. ELRA and ICCL.

[27] Hyunwoo Kim, Jihyeon Ryu, Jinho Lee, Hyungon Ryu, Kiran Praveen, Shyamala Prayaga, Kirit Thadaka, Will Jennings, Bardiya Sadeghi, Ashton Sharabiani, Yejin Choi, and Yev Meyer. Nemotron-Personas-Korea: Synthetic Personas Aligned to Real-World Distributions for Korea, April 2026.

[28] Hyunwoo Ko, Guijin Son, and Dasol Choi. Understand, Solve and Translate: Bridging the Multilingual Mathematical Reasoning Gap. In David Ifeoluwa Adelani, Catherine Arnett, Duygu Ataman, Tyler A. Chang, Hila Gonen, Rahul Raja, Fabian Schmidt, David Stap, and Jiayi Wang, editors, *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 78–95, Suzhuo, China, November 2025. Association for Computational Linguistics.

[29] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints. `https://arxiv.org/abs/2212.05055`, 2023.

[30] Michael S. Lee, Yash Maurya, Drew Rein, Bert Herring, Jonathan Nguyen, Kyungho Song, Udari Madhushani Sehwag, Jiyeon Cho, Kaustubh Deshpande, Yeongkyun Jang, Jiyeon Joo, Minn Seok Choi, Evi Fuelle, Christina Q. Knight, Joseph Brandifino, and Max Fenkell. ROK-FORTRESS: Measuring the Effect of Geopolitical Transcreation for National Security and Public Safety, 2026.

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

[32] Yucheng Li, Huiqiang Jiang, Yang Xu, Jianxin Yang, Yi Zhang, Yizhong Cao, Yuhao Shen, Fan Zhou, Rui Men, Jianwei Zhang, An Yang, Bowen Yu, Bo Zheng, Fei Huang, Junyang Lin, Dayiheng Liu, and Jingren Zhou. Breaking Entropy Bounds: Accelerating RL Training via MTP with Rejection Sampling. `https://arxiv.org/abs/2606.12370`, 2026.

[33] Jessy Lin, Vincent-Pierre Berges, Xilun Chen, Wen-Tau Yih, Gargi Ghosh, and Barlas Oğuz. Learning Facts at Scale with Active Reading. *arXiv preprint arXiv:2508.09494*, 2025.

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

[35] Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. Towards Robust Mathematical Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

[36] Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces, 2026.

[37] Yev Meyer and Dane Corneil. Nemotron-Personas-USA: Synthetic Personas Aligned to Real-World Distributions, June 2025.

[38] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. `https://arxiv.org/abs/2508.10925`, 2025.

[39] OpenAI. OpenAI MRCR: Long context multiple needle in a haystack benchmark. Hugging Face Datasets (snapshot), April 2025. Accessed: 2025-12-31.

[40] Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language, 2025.

[41] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow,

Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc

Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu,

Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Ben Wu, Jacek Karwowski, Davide Scaramuzza, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's Last Exam, 2025.

[42] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[43] Renjie Pi, Grace Lam, Mohammad Shoeybi, Pooya Jannaty, Bryan Catanzaro, and Wei Ping. On Data Engineering for Scaling LLM Terminal Capabilities, 2026.

[44] Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing Verifiable Instruction Following. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[45] 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. In *First Conference on Language Modeling*, 2024.

[46] Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. Reasoning to Learn from Latent Thoughts. *arXiv preprint arXiv:2503.18866*, 2025.

[47] Sheng Shen, Pete Walsh, Kurt Keutzer, Jesse Dodge, Matthew Peters, and Iz Beltagy. Staged Training for Transformer Language Models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[48] Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge, 2026.

[49] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics.

[50] StepFun Team. Step 3.5 Flash: Open Frontier-Level Intelligence with 11B Active Parameters. `https://arxiv.org/abs/2602.10604`, 2026.

[51] Artificial Analysis Team. Artificial Analysis Long Context Reasoning Benchmark (AA-LCR), 2025.

[52] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang,

Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi K3: Open Frontier Intelligence, 2026.

[53] Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, Xuanzhong Chen, Yida Zhao, Zhen Zhang, Zhengwei Tao, Zhongwang Zhang, Zile Qiao, Chenxi Wang, Donglei Yu, Gang Fu, Haiyang Shen, Jiayin Yang, Jun Lin, Junkai Zhang, Kui Zeng, Li Yang, Hailong Yin, Maojia Song, Ming Yan, Minpeng Liao, Peng Xia, Qian Xiao, Rui Min, Ruixue Ding, Runnan Fang, Shaowei Chen, Shen Huang, Shihang Wang, Shihao Cai, Weizhou Shen, Xiaobin Wang, Xin Guan, Xinyu Geng, Yingcheng Shi, Yuning Wu, Zhuo Chen, Zijian Li, and Yong Jiang. Tongyi DeepResearch Technical Report, 2026.

[54] Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. SciCode: A Research Coding Benchmark Curated by Scientists, 2024.

[55] Liang Wang, Nan Yang, Shaohan Huang, Li Dong, and Furu Wei. Thinking Augmented Pre-training, 2025.

[56] Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. Polymath: Evaluating mathematical reasoning in multilingual contexts. *Advances in Neural Information Processing Systems*, 38, 2026.

[57] 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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[58] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents, 2025.

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[60] Ryan Wong, Jiawei Wang, Junjie zhao, Li Chen, Yan Gao, Zhanglong, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and WANG KE. WideSearch: Benchmarking Agentic Broad Info-Seeking. In *The Fourteenth International Conference on Learning Representations*, 2026.

[61] Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. LLaMA Pro: Progressive LLaMA with Block Expansion, 2024.

[62] Ziyi Xia, Kun Luo, Hongjin Qian, Siqi Bao, and Zheng Liu. Open Data Synthesis for Deep Research. In *The Fourteenth International Conference on Learning Representations*, 2026.

[63] 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.

[64] John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[65] Bowen Ye, Rang Li, Qibin Yang, Yuanxin Liu, Linli Yao, Hanglong Lv, Zhihui Xie, Chenxin An, Lei Li, Lingpeng Kong, Qi Liu, Zhifang Sui, and Tong Yang. Claw-Eval: Towards Trustworthy Evaluation of Autonomous Agents, 2026.

[66] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019.

[67] 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.