From KMMLU-REDUX to PRO: A Professional Korean Benchmark Suite for LLM Evaluation

Seokhee Hong[1,⋆] **Sunkyoung Kim**[1,⋆] **Guijin Son**[2]
Soyeon Kim[1] **Yeonjung Hong**[1] **Jinsik Lee**[1]
[1]LG AI Research [2]OnelineAI

Abstract

The development of Large Language Models (LLMs) requires robust benchmarks that encompass not only academic domains but also industrial fields to effectively evaluate their applicability in real-world scenarios. In this paper, we introduce two Korean expert-level benchmarks. **KMMLU-REDUX**, reconstructed from the existing KMMLU (Son et al., 2024a), consists of questions from the Korean National Technical Qualification exams, with critical errors removed to enhance reliability. **KMMLU-PRO** is based on Korean National Professional Licensure exams to reflect professional knowledge in Korea. Our experiments demonstrate that these benchmarks comprehensively represent industrial knowledge in Korea. We release our dataset publicly available.[1]

1 Introduction

As LLMs continue to achieve strong performance across a wide range of subjects (OpenAI et al., 2024b; Deepmind, 2024; DeepSeek-AI et al., 2025a; Research et al., 2025b), the demand for comprehensive benchmarks has grown. MMLU (Hendrycks et al., 2021) is widely used for its broad coverage of general knowledge from elementary to college level. However, its publicly available online problems has raised concerns about reliability and potential data contamination (Gema et al., 2025; Vendrow et al., 2025; Zhao et al., 2024).

We identify similar issues in KMMLU (Son et al., 2024a), a widely used benchmark for evaluating Korean expert-level knowledge. The dataset was constructed by crawling websites that provide questions from various exams. We observe noisy samples, including problems that explicitly reveal the answer or non-existent reference, which can mislead performance evaluation. Additionally, we find evidence of contamination between the train and test splits, as well as with common web corpus such as FineWeb2 (Penedo et al., 2025).

Instead of collecting data from online sources directly, recent challenging benchmarks (Srivastava et al., 2023; Rein et al., 2024; Phan et al., 2025; Kazemi et al., 2025; Team et al., 2025b) have been constructed through problems and answers authored by human expert. Although this approach ensures high-quality, contamination-free benchmarks, it is costly to construct and maintain. The high construction cost hinders regular updates (White et al., 2025; Jain et al., 2025), leaving the benchmarks vulnerable to deprecation and contamination.

Furthermore, existing benchmarks primarily focus on academic knowledge and often overlook the practical applicability of models in industrial or professional contexts. As LLMs are increasingly adopted in industrial domains (Chkirbene et al., 2024), it becomes essential to assess whether they possess the necessary expertise to support tasks that require certified knowledge. For example, before deploying an LLM as a legal assistant, one must ensure that the model can reliably understand to meet professional certification standards.

We introduce two benchmarks to address the limitations above and incorporate professional knowledge to evaluate the practical applicability of LLMs. First, **KMMLU-REDUX**, a refined subset of KMMLU with 2,587 problems, is built through manual examination by authors to reduce errors and contamination within KMMLU. While the KMMLU contains wide range of problems, even in high-school level, KMMLU-REDUX only selects Korean National Technical Qualification (KNTQ) exams as sources. The exams require applicants to have either a bachelor's degree or at least nine

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

⋆ Authors equally contributed.

Email to: {seokhee.hong, sunkyoung.kim, jinsik.lee}@lgresearch.ai, guijin.son@onelineai.com

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19067–19096
November 4-9, 2025

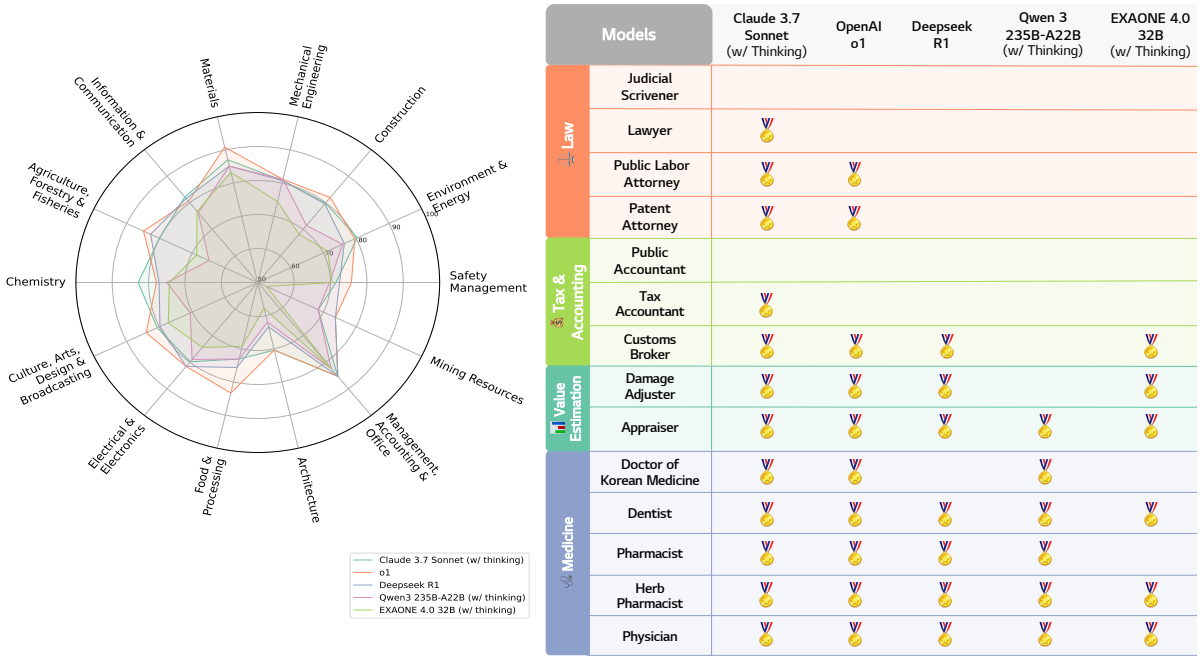

Figure 1: Performance of leading reasoning models developed by diverse groups on industrial knowledge for KMMLU-REDUX (**left**) and licensure exam pass status (indicated by 🏅) for KMMLU-PRO (**right**).

years experience in industrial field, thus making the benchmark more challenging.

Second, we build **KMMLU-PRO**, a new challenging benchmark, which consists of 2,822 problems from acquisition exams for Korean National Professional Licensure (KNPL), representing highly specialized professions in Korea. We include 14 professions (Table 1) from diverse domains. Unlike KMMLU that crawls websites, we collect data directly from the official source of each license. After that, human annotators manually examine it to avoid noises. KMMLU-PRO only includes exams held in the most recent year and would be updated annually with the latest exam to maintain long-term reliability and prevent contamination. See Appendix A for examples of each benchmark.

We conduct extensive evaluations of various LLMs on the two benchmarks. Our benchmarks, based on real-world exams, enable aligned analysis with industrial and professional qualifications, effectively revealing the practical strengths of each model. As shown in Figure 1, KMMLU-REDUX (left) covers a wide range of industrial domains, allowing it to evaluate the breadth of industrial knowledge. LLMs show robust performance in engineering domains but exhibit notable declines in specialized fields such as Mining & Resources and Architecture. On the other hand, KMMLU-PRO (right) focuses on professional licensure exams and thus assesses whether a model can pass the exams required for high-stakes professions in Korea. The results show that, in KMMLU-PRO, several state-of-the-art models perform strongly in the medicine domain, meeting the passing criteria of most licenses, yet nearly fail in law-related licenses.

Moreover, we observe the significant performance gaps between our datasets and merely translated datasets like MMMLU (OpenAI, 2024), especially in domains such as laws, where in-depth knowledge of specific countries is required (see Section 6.1). We argue that these findings underscore the practicality of our benchmarks for assessing the capabilities of models in professional fields within Korea.

To summarize, our contributions are as follows:

- We improve the previous benchmark, KMMLU, to construct the refined and compact version of the benchmark, **KMMLU-REDUX**, by correcting various errors.
- We introduce **KMMLU-PRO**, a new benchmark designed to evaluate high-level professional knowledge in Korea. By imitating real-world license acquisition systems, KMMLU-PRO assesses the industrial practicality across various professions in Korea.
- We comprehensively analyze the results of two benchmarks, highlighting the importance of benchmarks specialized in Korea-specific professional knowledge.

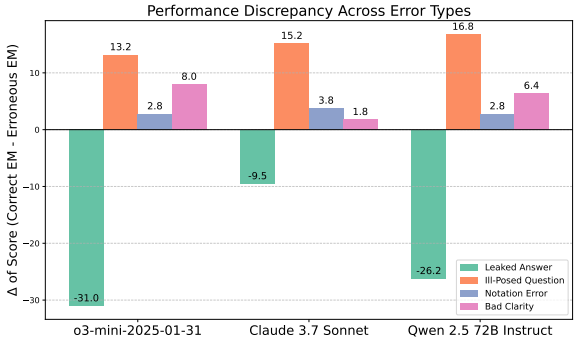

Figure 2: Performance differences in LLMs on the erroneous versus correct dataset. Leaked answer errors tend to overestimate model capabilities, while three other error types hinder LLMs to correctly predict true answers.

2 KMMLU-REDUX

We first revisit KMMLU (Son et al., 2024a) to examine its quality. Building on these insights, we construct KMMLU-REDUX, a cleaned and compact version of the KMMLU. We carefully denoise against the KMMLU and increase difficulty.

2.1 Revisiting KMMLU

KMMLU plays a significant role in the NLP community as a de facto standard for evaluating LLMs on Korea-specific expert knowledge (Yoo et al., 2024; Research et al., 2024; Team et al., 2025a). The dataset was constructed by crawling websites[2] where various exam questions are uploaded by people online, spanning from high school to professional qualification exams. Upon closer inspection, we identify several noises and limitations, which can be categorized into three types: 1) duplication issues, 2) dataset errors, and 3) contamination.

Duplication Issue As the KMMLU crawls problems from hundreds of exams in Korea, we observe multiple duplicated questions across related exams. Notably, duplicated samples occur not only within the test set but also between the training and test sets. Using the Longest Common Sequence (LCS) algorithm to investigate overlaps, we could find 5.36% duplication within the test set and a 5.46% contamination between train and test set.

Dataset Errors Following Gema et al. (2025), we investigate the extent to which various error types appear in the KMMLU test set and their potential impact on LLM performance. We identify four representative error types: leaked answers, ill-posed questions, poor clarity, and notation errors. We then annotate the entire test set using GPT-4o[3] (OpenAI et al., 2024a). In total, we find that 7.66% of the data contains one of the errors mentioned above. We describe the details of the investigation of dataset errors in Appendix B.

In Figure 2, we observe significant discrepancies in the LLM's performance between erroneous and correct samples. Notably, the performance of instances with leaked answers drops significantly when the LLMs are assessed on the clean dataset. Conversely, all models' scores increase on ill-posed questions, underscoring their inability to identify the correct answer for poorly formulated questions.

Contamination The KMMLU was primarily sourced online, making them highly susceptible to contamination from web-crawled training corpora. When applying n-gram contamination detection (Lambert et al., 2025; Grattafiori et al., 2024) to the Korean subset of FineWeb2 (Penedo et al., 2025) and the KMMLU dataset, 1.88% of the data were flagged as contaminated.

2.2 Dataset Construction

KMMLU consists of approximately $35k$ examples, making evaluation resource-intensive. To reduce this burden, we first restrict the scope to a subset of high-difficulty exams (Section 2.2.1). Furthermore, to ensure the reliability of the benchmark, we manually conduct a thorough examination of the dataset to eliminate errors (Section 2.2.2).

2.2.1 Filtering Non-Challenging Problems

Since KMMLU includes a variety of exams in Korea, spanning from high school to professional certification exams, we filter out the easier exams to build a more challenging benchmark. Specifically, we choose Korean National Technical Qualification (KNTQ) exams, which are primarily designed to assess practical technical competencies required in industrial field.

The qualifications require applicants to have either a bachelor's degree or at least nine years of professional experience. We adopt a collection of 100 KNTQ exams across the 14 domains in total. We only include the most recent exam for each qualification, thereby avoiding outdated knowledge being evaluated[4].

[2] https://www.kinz.kr/

[3] gpt-4o-2024-11-20

[4] We have compiled a list of all KNTQs exams alongside their most recent exam dates in Appendix C. Our aim is to make these available to LLM researchers and developers to help prevent data contamination.

Domain	Names of KNPLs	U.S. Equivalent	# of Instances
Law	Certified Judicial Scrivener	Paralegal, Legal Document Assistant, or Notary Public (no direct equivalent)	198
	Lawyer (Kim et al., 2024b)	Attorney-at-Law	150
	Certified Public Labor Attorney	Labor & Employment Lawyer (requires J.D. and bar admission; no separate certification in the U.S.)	239
	Certified Patent Attorney	Patent Attorney (JD + USPTO registration required)	109
Tax & Accounting	Certified Public Accountant (CPA)	Certified Public Accountant (CPA) – Exact Equivalent	208
	Certified Tax Accountant	Enrolled Agent (IRS) or CPA with Tax Specialization	238
	Certified Customs Broker	U.S. Customs Broker (licensed by U.S. Customs and Border Protection - CBP)	159
Value Estimation	Certified Damage Adjuster (CDA)	Claims Adjuster / Insurance Adjuster (state-licensed)	120
	Certified Appraiser	Certified Real Estate Appraiser (licensed at the state level)	196
Medicine	Doctor of Korean Medicine	Licensed Acupuncturist (L.Ac.) or Doctor of Acupuncture and Oriental Medicine (D.A.O.M.)	288
	Dentist (Kweon et al., 2024)	Doctor of Dental Surgery (D.D.S.) / Doctor of Dental Medicine (D.M.D.)	252
	Pharmacist (Kweon et al., 2024)	Doctor of Pharmacy (Pharm.D.)	271
	Herb Pharmacist	Herbalist (non-licensed or CAM-certified depending on state)	244
	Physician (Kweon et al., 2024)	Medical Doctor (M.D./D.O.)	150
	Total		**2822**

Table 1: The list of National Professional Licenses (NPLs) used for KMMLU-PRO and their corresponding statistics. The names of NPLs are translated from those in Korea and we also report equivalent licences in U.S. We use KorMedMCQA (Kweon et al., 2024) for three licenses in the Medical category, and KBL (Kim et al., 2024b) for the bar exam of lawyer.

To further discriminate simple problems, we leverage the performances of LLMs on the data (Wang et al., 2024; Zellers et al., 2019; Lee et al., 2023). By employing seven smaller LLMs[5], we mark a data as *easy* if four or more models correctly predict its answer. Through this process, we remove 38.6% of the dataset.

2.2.2 Denoising

To remove noises, we follow the processes described in Section 2.1. Specifically, we first manually review the dataset to minimize errors. Next, we perform decontamination to prevent potential data leakage from pre-training corpora. Additionally, we detect inner duplication and against the training and test sets in KMMLU, thus finally remove all duplicates.

2.2.3 Final Statistics

For KMMLU-REDUX, we have collected 2,587 problems from 100 KNTQ exams. Among these, 596 problems are from exams that require over nine years of professional experience to acquire the qualification. To categorize the dataset into 14 domains, we follow the Korean Standard Industrial Classification (KSIC) published by Statistics Korea[6] as the qualification system is primarily designed to align with industrial fields. Figure 7 in Appendix C illustrates the distribution of domain of KMMLU-REDUX.

3 KMMLU-PRO

A major challenge in building benchmarks from online sources is data contamination (Zhao et al., 2024; Jain et al., 2025; Roberts et al., 2024). While some studies address this by having experts manually create problems (Srivastava et al., 2023; Rein et al., 2024; Phan et al., 2025; Kazemi et al., 2025; Team et al., 2025b), this approach is costly and time-consuming. As an alternative, recent work explores periodic releases of fresh subsets. (White et al., 2025; Jain et al., 2025).

Motivated by these approaches, we focus on the Korean National Professional Licensure (KNPL) exams. These are high-stakes exams administered annually that pose a significant challenge. Unlike benchmarks crafted by a small set of experts (Rein et al., 2024; Phan et al., 2025), our approach leverages the well-established curricula of professional licensing systems, designed to assess real-world professional knowledge.

3.1 Korean National Professional Licensures

We choose KNPL exams for main source of KMMLU-PRO. KNPL exams target high-level professionals, such as lawyers, accountants, or physicians, requiring advanced knowledge, critical reasoning, and ethical judgment. Among them, we select 14 KNPLs representing highly specialized and regulated professions in Korea (See Table 1 for the list of KNPLs and their equivalent licensure in U.S.). These licenses are legally mandated creden-

[5]Llama 3.2 3B (Meta, 2024c), Qwen 2.5 3B (Qwen et al., 2025), Gemma 3 4B IT (Team, 2025a), Kanana Nano 2.1B Instruct (Team et al., 2025a), EXAONE 3.5 2.4B (Research et al., 2024), DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025a) EXAONE Deep 2.4B (Research et al., 2025b), and Ko-R1-7B-v2.1 (OneLineAI, 2025).

[6]https://www.kostat.go.kr/

tials required to practice in their respective domains. As such, they serve as institutionalized gateways to high-status occupations with significant entry barriers. Our evaluation simulates real-world assessment standards by incorporating official exam pass criteria, aligning model performance with human standards.

3.2 Dataset Collection and Annotation

In Korea, the government releases and manages the questions for KNPL acquisition exams. We directly download the PDF files from the government's websites for each license and use GPT-4o (OpenAI et al., 2024a) for OCR parsing. As our dataset sources from the official PDFs, we can enhance the quality of the dataset, avoiding potential errors when collecting from online text (Team et al., 2025b). Since GPT-4o has difficulty handling tables and low-resolution PDFs, we employ human annotators to review parsed questions. When a problem contains an image, the annotators convert it into text that conveys the same meaning of the image, if possible. Notably, our process remains relatively cost-efficient as it requires human annotators solely for error reviewing tasks, not full annotation. We demonstrate detailed annotation process in Appendix D

We follow previous works, releasing a new set of questions periodically (White et al., 2025; Jain et al., 2025). Since the exams are conducted annually, we commit to collect and release questions from the exams held just before the current year.

3.3 Decontamination

We also adopt the same process outlined in Section 2.2.2 to ensure KMMLU-Pro is free from contamination. When conducting n-gram match between KMMLU-Pro, FineWeb2, and the training and validation sets of KMMLU, we did not find any contaminated examples. As a result, we can retain all 2,822 data points in the KMMLU-Pro, maintaining its contamination-free integrity.

4 Experiments

We select a diverse set of baseline models varying in size, multilingual capability, and reasoning ability. By default, we apply a zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022) prompt written in Korean. However, in early experiments, we observed that some reasoning models perform worse when prompted in Korean. Therefore, we evaluate those models using both Korean and English prompts and report the scores with the highest average score across KMMLU-Redux and KMMLU-Pro[7]. We use greedy decoding for non-reasoning models, while for reasoning models, we adopt the temperature of 0.6 and top-p (Holtzman et al., 2020) of 0.95. For more details, see Appendix E.

Evaluating open-weight models was conducted over four days using sixteen NVIDIA A100 GPUs. For closed models, we accessed them via API calls, which incurred a total cost of approximately $4,000.

Metrics In addition to accuracy, our primary evaluation metric, we also report the number of licensure exams each LLM passes in KMMLU-Pro. To better align with human evaluation standards, our procedure is designed to mirror the official certification standards. However, for licenses that rely heavily on image-based questions, full replication is not possible. See Appendix E.3 for details.

5 Results

5.1 Main Results

Table 2 presents the overall performance on KMMLU-Redux and KMMLU-Pro. The o1 model achieves the highest average accuracy (79.55), followed by Claude 3.7 with Thinking (78.49). Among open-weight models, DeepSeek's R1 even outperforms many closed models. Models equipped with reasoning capabilities consistently perform better than their non-reasoning models, such as the Qwen3 series and EXAONE 4.0.

Beyond accuracy, we evaluate each model's ability to pass the licensure exams under the official criteria for KMMLU-Pro. Specifically, in most licenses, a model must score at least 40% in each subject and achieve an overall average of 60% to pass. Claude 3.7 with Thinking succeeds in 12 out of 14 KNPL licensure exams, the highest among all evaluated models. In contrast, although the o1 model achieves higher accuracy, it qualifies for fewer licenses than Claude 3.7 with Thinking. This highlights the importance of balanced competence across subjects, as required in real-world certification exams.

[7]For further details, please see Section 6.4.

	KMMLU-Redux	KMMLU-Pro		
	Acc	Acc	# of passed KNPLs	Avg. Acc (micro)
Open-weight Models				
Aya Expanse 32B (Dang et al., 2024)	33.05	31.26	0/14	32.12
Gemma 3 12B IT (Team, 2025a)	46.70	45.82	2/14	46.24
Phi-4 (14B) (Abdin et al., 2024)	49.75	45.32	1/14	47.44
Mistral Small 3.1 Instruct (24B) (Mistral, 2025)	52.92	49.49	3/14	51.13
Gemma 3 27B IT (Team, 2025a)	54.04	51.03	2/14	52.47
Llama 3.3 70B Instruct (Meta, 2024a)	56.17	53.24	3/14	54.64
Qwen3-14B (Yang et al., 2025)	57.25	53.02	3/14	55.04
Qwen3-30B-A3B (Yang et al., 2025)	58.41	52.33	3/14	55.24
C4AI Command A (111B) (Cohere, 2025)	62.93	57.48	3/14	60.07
Qwen3-32B (Yang et al., 2025)	64.98	58.86	3/14	61.79
EXAONE 4.0 32B (Research et al., 2025a)	64.79	60.01	3/14	62.30
Llama-4-Scout-17B-16E-Instruct (Meta, 2025)	67.49	58.14	4/14	62.61
Qwen3-30B-A3B (w/ thinking) (Yang et al., 2025)	65.25	60.52	3/14	62.78
Qwen3-14B (w/ thinking) (Yang et al., 2025)	65.71	60.18	2/14	62.82
DeepSeek V3 (671B) (DeepSeek-AI et al., 2025b)	65.64	60.77	4/14	63.10
Qwen3-32B (w/ thinking) (Yang et al., 2025)	68.77	61.14	3/14	64.79
QwQ 32B (Team, 2025b)	67.34	63.94	5/14	65.57
Qwen3-235B-A22B (Yang et al., 2025)	69.54	62.12	4/14	65.67
EXAONE 4.0 32B (w/ thinking) (Research et al., 2025a)	72.71	67.67	6/14	70.08
Qwen3-235B-A22B (w/ thinking) (Yang et al., 2025)	74.49	68.22	6/14	71.22
Llama-4-Maverick-17B-128E-Instruct (Meta, 2025)	77.58	68.10	4/14	72.63
DeepSeek R1 (671B) (DeepSeek-AI et al., 2025a)	78.51	71.33	7/14	74.76
Closed Models				
GPT-4.1 mini (2025-04-14) (OpenAI, 2025a)	67.03	62.18	4/14	64.50
o3-mini (2025-01-31) (OpenAI, 2025c)	67.84	62.05	3/14	64.82
Grok-3-mini-beta (xAI, 2025)	71.47	65.08	5/14	68.14
Grok-3-beta (xAI, 2025)	72.90	68.37	7/14	70.54
o4-mini (2025-04-16) (OpenAI, 2025b)	75.80	69.65	6/14	72.59
GPT-4.1 (2025-04-14) (OpenAI, 2025a)	75.86	72.99	10/14	74.36
Claude 3.7 Sonnet (Anthropic, 2025)	76.88	74.52	10/14	75.65
o3 (OpenAI, 2025b)	79.92	73.60	9/14	76.62
Claude 3.7 Sonnet (w/ thinking) (Anthropic, 2025)	79.36	77.70	12/14	78.49
o1 (OpenAI et al., 2024b)	81.14	78.09	10/14	79.55

Table 2: The main evaluation results of KMMLU-REDUX and KMMLU-PRO benchmarks on various LLMs. The gray-shaded models stand for reasoning models. The results of models with size < 10B are presented in Table 7 in Appendix F.1.

5.2 Performance Across Industrial Domains in KMMLU-REDUX

Through KMMLU-REDUX, we assess the technical competencies of models across diverse industrial fields. As shown in Figure 1 (left), models demonstrate varying levels of performance across these fields. For example, leading models with reasoning capabilities struggle in domains such as Safety Management, Mining Resources, and Architecture, achieving under 80% accuracy, which highlights persistent challenges in underrepresented or highly specialized fields. In contrast, the models obtain scores exceeding 80% in Materials and Management, Accounting & Office areas. The full results for all models across 14 domains are presented in Appendix F.2.

5.3 Professional Licensure Exam Performance on KMMLU-PRO

We provide a breakdown of the KMMLU-PRO results by analyzing which licensure exams are most frequently qualified by LLMs. Figure 1 (right) shows the pass rates of LLMs across 14 KNPLs. While many models pass the exams in the medicine domain, most fail in Law and Tax & Accounting, with only DeepSeek R1 and EXAONE 4.0 32B passing the Customs Broker exam among open-weight models. Notably, no models pass the Judicial Scrivener or Public Accountant exams. This trend becomes even more evident in the full results across a broader range of LLMs; the only licenses

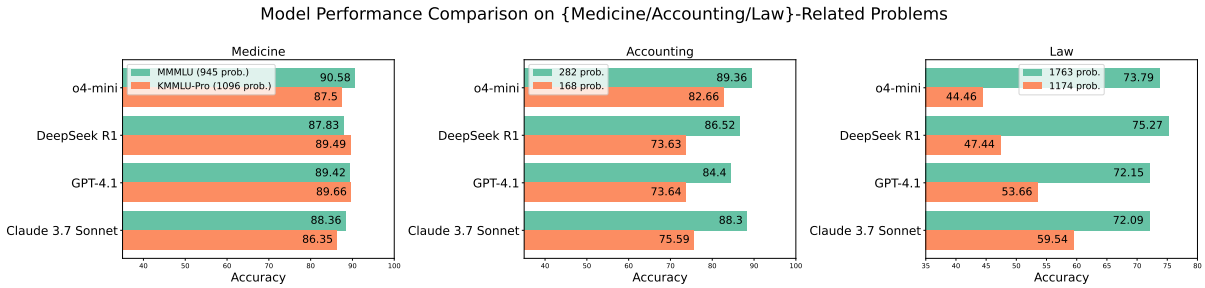

Figure 3: Performance of four LLMs on {Medical**(left)**, Accounting**(center)**, Law**(right)**}-relevant subsets from the MMMLU (Korean) (OpenAI, 2024) and KMMLU-PRO. While the discrepancies in scores are narrow in the medicine domain, they are wider in law-related problems, emphasizing the need for datasets that reflecting real professional knowledge in Korea.

that relatively smaller models (<20B) are able to pass in the medicine domain (see Table 9 in Appendix F.3).

Moreover, many LLMs fail in the licensure exams even when scoring above 60%; for example, o3-mini, Qwen3-235B-A22B, and Llama-4-Maverick score over 85% on the Pharmacist exam but still fail to qualify due to not meeting the threshold of law-related subject in the exam. These cases highlight the difficulty of acquiring region-specific domain knowledge, particularly in legal subjects governed by Korean law.

6 Analysis

6.1 The Importance of Locally Adapted Benchmarks

To highlight the importance of evaluation grounded in local context (Plaza et al., 2024; Singh et al., 2025), we compare category-level performance between benchmarks translated from English and KMMLU-PRO. Specifically, we focus on subjects related to law, accounting, and medicine, [8] selecting relevant subjects from the Korean subset of MMMLU (OpenAI, 2024), as well as KMMLU-PRO.

As shown in Figure 3, the performance gap is relatively small in categories such as medicine, where domain knowledge is largely consistent across countries and cultures. In contrast, categories such as law, where substantial differences in content are expected, show a significantly larger gap. This suggests that MMMLU, which relies on direct translation of law questions based on U.S. standards, cannot adequately represent knowledge of Korean law. These findings highlight the importance of our dataset, which reflects authentic professional knowledge specific to the Korean context.

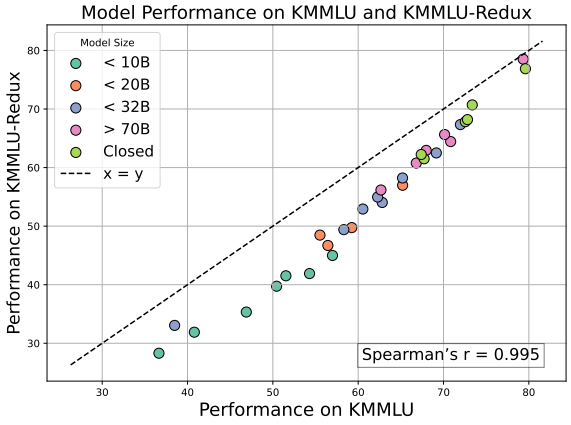

Figure 4: Performance of LLMs on KMMLU and KMMLU-REDUX. A high ρ value indicates a strong correlation between the results of the two benchmarks.

6.2 KMMLU vs KMMLU-REDUX

Figure 4 illustrates the performance of LLMs on KMMLU and KMMLU-REDUX. The LLMs' performances on KMMLU-REDUX is lower than on KMMLU, due to our filtration process which aims to retain only challenging problems from KMMLU (see Section 2.2.1). Despite this decrease, there is a near-perfect monotonic association between the results, with a Spearman's rank correlation coefficient (ρ) of 0.995, suggesting they are highly correlated.

6.3 Impact of Reasoning Budget

Recent studies have shown that increasing reasoning efforts can enhance model performance (Muennighoff et al., 2025; Anthropic, 2025; Yang et al., 2025). To examine how reasoning *budget* affects

[8] {*professional_law*, *jurisprudence*, *international_law*} for Law. {*professional_accounting*} for Accounting. {*professional_medicine*, *clinical_knowledge*, *college_medicine*, *medical_genetics*, *anatomy*} for Medicine.

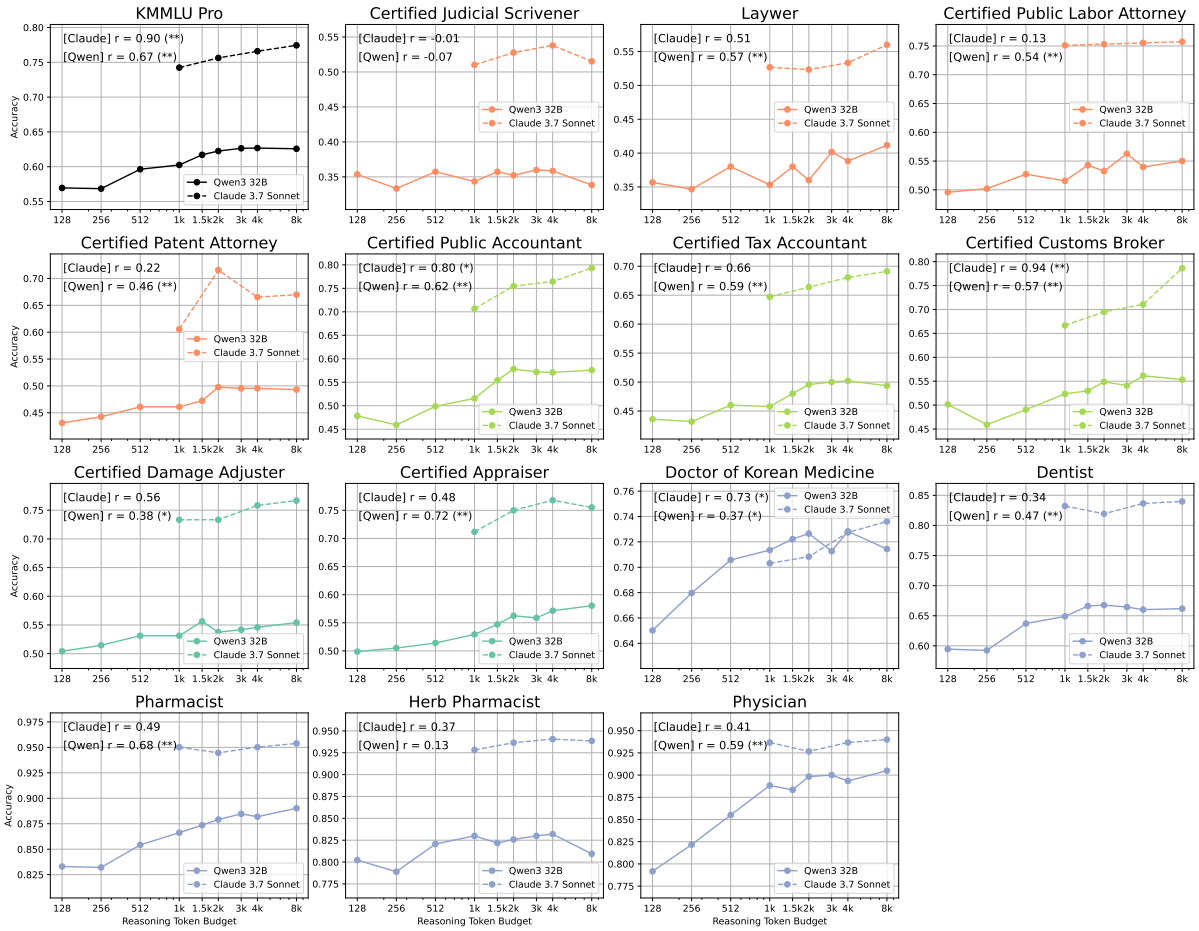

Figure 5: Reasoning budget results of Qwen3-32B and Claude 3.7 Sonnet on KMMLU-PRO. r indicates the Pearson Correlation coefficient value between the reasoning token budget and the accuracy. The responses are sampled multiple times for each thinking budget setting; $n = 4$ and $n = 2$, respectively, for Qwen and Claude. * and ** denote statistical significance, indicating p-value <0.05 and <0.01, respectively.

performance on each licenses in KMMLU-PRO, we conduct an experiment varying the number of tokens allocated to the reasoning path. We adopt Qwen3-32B (Yang et al., 2025) and Claude 3.7 Sonnet (Anthropic, 2025)[9]. For each budget $b \in B$, we generate n different responses per question and average the scores. We set $n = 4$ for Qwen3, but $n = 2$ for Claude due to the generation cost.

As shown in Figure 5, we observe a positive correlation between the reasoning budget and the overall score of KMMLU-PRO for both models. However, this trend does not hold uniformly across all licenses. For example, both models show trivial gains on the Judicial Scrivener and Herb Pharmacist licenses, indicating that more reasoning does not always boost performance.

We further analyze the effect of enabling reasoning itself. To be specific, we compare the scores of reasoning models with their corresponding non-reasoning models on performance in KMMLU-PRO licenses, observing how enabling reasoning impacts results. As shown in Table 10 in Appendix F.4, reasoning models exhibit different trends across licenses. For example, in the Judicial Scrivener exam, most reasoning models do not show significant performance gains over their non-reasoning counterparts, except for Qwen3-1.7B, which aligns with the results of Figure 5. In contrast, enabling reasoning boosts performance for many models on CPA exams, demonstrating the impact of reasoning on calculation-intensive tasks.

6.4 Impact of Prompt Language

Prompt language can significantly influence model behavior, raising concerns about consistency in multilingual settings (Wang et al., 2025; Zhang et al., 2025; Lai and Nissim, 2024). Since all

[9]We select the models because they either natively support the "thinking budget" or have reported experimental results on thinking budgets in their technical report.

	KMMLU-Redux			KMMLU-Pro		
	English	Korean	*diff* (%)	English	Korean	*diff* (%)
Qwen3-32B (w/ thinking)	68.77	69.08	+0.5%	61.14	60.66	-0.8%
o4-mini (2025-04-16)	75.80	76.17	+0.5%	69.65	69.10	-0.8%
Qwen3-235B-A22B (w/ thinking)	74.49	75.11	+0.8%	68.22	66.98	-1.8%
Grok-3-mini-beta	71.47	70.85	-0.9%	65.08	64.89	-0.3%
Qwen3-14B (w/ thinking)	65.71	65.40	-0.5%	60.18	59.48	-1.2%
EXAONE Deep 32B	58.33	56.17	-3.7%	52.33	52.19	-0.3%
DeepSeek R1 (671B)	78.51	75.38	-4.0%	71.33	70.62	-1.0%
QwQ 32B	67.34	62.66	-6.9%	63.94	59.95	-6.2%
EXAONE Deep 7.8B	44.99	40.82	-9.3%	41.53	38.98	-6.1%
Llama-4-Maverick-17B-128E-Instruct	77.58	72.52	-7.0%	68.10	57.15	-16.1%
Llama-4-Scout-17B-16E-Instruct	67.49	45.03	-33.3%	58.14	28.74	-50.6%

Table 3: Comparison results between English and Korean prompts of models whose main results are reported on English prompts. The *diff* values are the relative difference in scores between two prompts. The specific prompts are detailed in Appendix E.1.

questions in our datasets are written in Korean, using Korean prompts is a natural choice. However, we observe that some models perform worse when prompted in Korean. Table 3 presents the performance difference between English and Korean prompts. The Llama-4 model series exhibits the most substantial drop in performance[10], while closed models such as Grok-3-mini-beta and o4-mini show minimal change.

7 Related Works

Reliability Issues of Benchmarks Recent works (Gema et al., 2025; Vendrow et al., 2025) have raised concerns about the reliability of LLM benchmarks due to dataset noise and contamination. MMLU-Redux (Gema et al., 2025) improved evaluation quality through systematic human reannotation, while GSM8K-Platinum (Vendrow et al., 2025) refined arithmetic benchmarks via automated and manual error detection. MMLU-CF (Zhao et al., 2024) prevents both unintentional and malicious contamination via sourcing diverse domains and question rewriting. LiveBench (White et al., 2025) and LiveCodeBench (Jain et al., 2025) adopted dynamic evaluation protocols with temporal cutoffs to prevent future leakage.

Professional Benchmark With the rapid advancement of LLMs, more challenging benchmarks have become essential. GPQA (Rein et al., 2024) and SuperGPQA (Team et al., 2025b) assess graduate-level knowledge; MMLU-Pro (Wang et al., 2024) extends MMLU by increasing the share of college-level questions and expanding answer choices. Humanity's Last Exam (Phan et al., 2025) introduces a frontier benchmark composed of manually authored, research-level questions.

Korean Benchmark While prior benchmarks focus primarily on English, recent efforts have produced Korean-specific evaluations (Son et al., 2024b; Kim et al., 2024a). Some rely on translated datasets (Park et al., 2024; Kim et al., 2025; OpenAI, 2024; Singh et al., 2024), often with human post-editing, but these lack regional context, institutional norms, and domain-specific fluency. In contrast, native Korean benchmarks such as KMMLU (Son et al., 2024a), KorMedMCQA (Kweon et al., 2024), and KBL (Kim et al., 2024b) address cultural and linguistic specificity. However, KMMLU (Son et al., 2024a) suffers from quality issues, including leaked answers, and KorMedMCQA (Kweon et al., 2024) and KBL (Kim et al., 2024b) are limited to narrow domains.

In this work, we introduce KMMLU-Redux and KMMLU-Pro, two contamination-free, industry grade benchmarks, providing a practical assessment of LLM capabilities in Korean industries.

8 Conclusion

We present two benchmarks constructed from real-world professional licensing exams, designed to reflect industrial domain knowledge and practical application standards. To ensure reliability, we identify and eliminate various sources of errors. Through extensive experiments, we evaluate the professional knowledge capabilities of LLMs across a wide range of domains. Our analysis further identifies key factors that influence performance, including region-specific knowledge, reasoning budget, and prompt language. We hope this work provides a foundation for more rigorous evaluation and continued advancement of real-world competence in language models.

Limitations

Our benchmarks are limited to text-only and multiple-choice questions for text-only LLMs. It restricts its coverage of real-world licensure exams. Many real-word professional qualification exams include non-textual modalities or require constructed responses such as essay. Our benchmark cannot fully assess all aspects of professional competence or reasoning required in such exams. Expanding to multimodal inputs and open-ended question formats is an important direction for future work.

[10]The Llama-4 models did not follow the prompt but ended their response with "*The best answer is*", even with the Korean prompt.

Ethical Statements

All data used in our benchmarks are either publicly available or collected from official licensing materials released by government or professional institutions. For quality control, we hired human annotators to review parsed questions from PDF; they were compensated over the mininum wage in Korea. Our benchmarks would be released under CC-BY-NC-ND 4.0 license.

References

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

Anthropic. 2025. Claude 3.7 sonnet and claude code.

Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. 2024. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234.

Cohere. 2025. Model card for c4ai command a.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Google Deepmind. 2024. Introducing gemini 2.0: our new ai model for the agentic era.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, 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, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

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

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. 2025. Big-bench extra hard. *Preprint*, arXiv:2502.19187.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Hyeonwoo Kim, Dahyun Kim, Jihoo Kim, Sukyung Lee, Yungi Kim, and Chanjun Park. 2025. Open ko-llm leaderboard2: Bridging foundational and practical evaluation for korean llms. *Preprint*, arXiv:2410.12445.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024b. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations. *Preprint*, arXiv:2403.01469.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. 2023. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6692–6712, Toronto, Canada. Association for Computational Linguistics.

Meta. 2024a. The future of ai: Built with llama.

Meta. 2024b. Introducing llama 3.1: Our most capable models to date.

Meta. 2024c. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Mistral. 2025. Mistral small 3.1.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

naver hyperclovax. 2025. Model card for hyperclovax-seed-text-instruct-1.5b.

OneLineAI. 2025. Ko-r1-7b-v2.1. `https://huggingface.co/OLAIR/ko-r1-7b-v2.1`. Accessed: 26 March 2025.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani,

Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai

Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu).

OpenAI. 2025a. Introducing gpt-4.1 in the api.

OpenAI. 2025b. Introducing openai o3 and o4-mini.

OpenAI. 2025c. Openai o3-mini.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, SeongHwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024. Open Ko-LLM leaderboard: Evaluating large language models in Korean with Ko-h5 benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3220–3234, Bangkok, Thailand. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

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, Tung Nguyen, Daron Anderson, Imad Ali Shah, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Jaeho Lee, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, Robert Gerbicz, John-Clark Levin, Serguei Popov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Mstyslav Kazakov, Geoff Galgon, Johannes Schmitt, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Antrell Cheatom, Zachary Giboney, Gashaw M. Goshu, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, Jennifer Zampese, John B. Wydallis, Ryan G. Hoerr, Mark Nandor, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Jungbae Nam, Edwin Taylor, Jun Jin, Gautier Abou Loume, Hangrui Cao, Alexis C Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Aras Bacho, Lianghui Li, Sumeet Motwani, Christian Schroeder de Witt, Alexei Kopylov, Johannes Veith, Eric Singer, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Ameya Prabhu, Longke Tang, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Joshua Robinson, Aleksandar Mikov, Julien Guillod, Yuqi Li, Ben Pageler, Joshua Vendrow, Vladyslav Kuchkin, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Andrew Gritsevskiy, Dakotah Martinez, Nick Crispino, Dimitri Zvonkine, Natanael Wildner Fraga, Saeed Soori, Ori Press, Henry Tang, Julian Salazar, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, T. Ryan Rogers, Wenjin Zhang, Ross Finocchio, Bikun Li, Jinzhou Yang, Arun Rao, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Ariel Ghislain Kemogne Kamdoum, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P Coppola, Haline Heidinger, Rafael Sayous, Stefan Ivanov, Joseph M Cavanagh, Jiawei Shen, Joseph Marvin Imperial, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Brecht Verbeken, Kelsey Van den Houte, Lynn Van Der Sypt, David Noever, Lisa Schut, Ilia Sucholutsky, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Shankar Sivarajan, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Jennifer Sandlin, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Felipe Meneguitti Dias, Tobias Kreiman, Kaivalya Rawal, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Sergey Ivanov, Rafał Poświata, Chenguang Wang, Daofeng Li, Donato Crisostomi,

Ali Dehghan, Andrea Achilleos, John Arnold Ambay, Benjamin Myklebust, Archan Sen, David Perrella, Nurdin Kaparov, Mark H Inlow, Allen Zang, Kalyan Ramakrishnan, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Dan Bar Hava, Aleksey Kuchkin, Robert Lauff, David Holmes, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Daniel Pyda, Zakayo Kazibwe, Mukhwinder Singh, Don Clarke, Dae Hyun Kim, Sara Fish, Veit Elser, Victor Efren Guadarrama Vilchis, Immo Klose, Christoph Demian, Ujjwala Anantheswaran, Adam Zweiger, Guglielmo Albani, Jeffery Li, Nicolas Daans, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Ziqiao Ma, Christian Stump, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Marco Piccardo, Niv Cohen, Virendra Singh, Josef Tkadlec, Paul Rosu, Alan Goldfarb, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stade, Declan Grabb, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, Abhishek Shukla, Hossam Elgnainy, Yan Carlos Leyva Labrador, Hao He, Ling Zhang, Alan Givré, Hew Wolff, Gözdenur Demir, Muhammad Fayez Aziz, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Elliott Thornley, Robin Zhang, Jiayi Pan, Antonio Terpin, Niklas Muennighoff, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Jainam Shah, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Andrew Ho, Shaul Barkan, Jiaqi Wang, Martin Stehberger, Egor Kretov, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Zaki Hossain, Ido Akov, 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, Micah Carroll, Orr Paradise, Jan Hendrik Kirchner, Stefan Steinerberger, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Paolo Giordano, Philipp Petersen, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Antonella Pinto, Shreyas Verma, Prashant Joshi, Eli Meril, Zheng-Xin Yong, Allison Tee, Jérémy Andréoletti, Orion Weller, Raghav Singhal, Gang Zhang, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Hamid Mostaghimi, Kunvar Thaman, Qijia Chen, Tran Quoc Khánh, Jacob Loader, Stefano Cavalleri, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Jonathan Roberts, William Alley, Kunyang Sun, Ryan Stendall, Max Lamparth, Anka Reuel, Ting Wang, Hanmeng Xu, Pablo Hernández-Cámara, Freddie Martin, Thomas Preu, Tomek Korbak, Marcus Abramovitch, Dominic Williamson, Ida Bosio, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Maria Inês S. Nunes, Yibo Jiang, M Saiful Bari, Peyman Kassani, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Guillaume Douville, Daniel Tordera, George Balabanian, Earth Anderson, Lynna Kvistad, Alejandro José Moyano, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Isaac C. McAlister, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Ronald Clark, Sherwin Abdoli, Tim Santens, Harrison K Wang, Evan Chen, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Niels Mündler, Avi Semler, Emma Rodman, Jacob Drori, Carl J Fossum, Luk Gloor, Milind Jagota, Ronak Pradeep, Honglu Fan, Tej Shah, 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, Siranut Usawasutsakorn, Mohammadreza Mofayezi, Alexander Piperski, Marc Carauleanu, David K. Zhang, Kostiantyn Dobarskyi, Dylan Ler, Roman Leventov, Ignat Soroko, Thorben Jansen, Scott Creighton, Pascal Lauer, Joshua Duersch, Vage Taamazyan, Dario Bezzi, Wiktor Morak, Wenjie Ma, William Held, Tran Đuc Huy, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Hossein Shahrtash, Edson Oliveira, Joseph W. Jackson, Daniel Espinosa Gonzalez, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Emilien Duc, Bita Golshani, David Stap, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Lukas Lewark, Miguel Orbegozo Rodriguez, Mátyás Vincze, Dustin Wehr, Colin Tang, Shaun Phillips, Fortuna Samuele, Jiang Muzhen, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Claire Sparrow, Rayner Hernandez Perez, 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, Samuel Albanie, Will Cai, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Jasdeep Sidhu, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Brian Weber, Harsh Kumar, Tong Jiang, Arunim Agarwal, Chiara Ceconello, Warren S. Vaz, Chao Zhuang, Haon Park, Andrew R. Tawfeek, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Shreen Gul, Gunjan Chhablani, Zhehang Du, 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, Florencia de la Rosa, Xiuyu Li, Guillaume Malod, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Luca Arnaboldi, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Pierre Clavier, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Denis Peskoff, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu,

Liu, Olle Häggström, Emil Verkama, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Yiyang Fan, Gabriel Poesia Reis e Silva, Linwei Xin, Yosi Kratish, Jakub Łucki, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Justin Xu, Kevin Joseph Scaria, Freddie Vargus, Farzad Habibi, Long, Lian, Emanuele Rodolà, Jules Robins, Vincent Cheng, Tony Fruhauff, Brad Raynor, Hao Qi, Xi Jiang, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, 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, Sarah Hoback, Rodrigo De Oliveira Pena, Glen Sherman, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Sam Ali, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Sk Md Salauddin, Murat Islam, Juan Gonzalez, Josh Ducey, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I. M. J. McInnis, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Javier Gimenez, Roselynn Grace Montecillo, Russell Campbell, Asankhaya Sharma, Khalida Meer, Xavier Alapont, Deepakkumar Patil, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Sergei Bogdanov, Sören Möller, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Innocent Enyekwe, Ragavendran P V, Zienab EL-Wasif, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Song Bian, John Lai, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Alex Hoover, Joseph McGowan, Tejal Patwardhan, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation? *Preprint*, arXiv:2406.17789.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

LG AI Research, :, 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. 2025a. Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes. *Preprint*, arXiv:2507.11407.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Sihoon Yang, Heuiyeen Yeen, and Hyeongu Yun. 2024. Exaone 3.5: Series of large language models for real-world use cases. *Preprint*, arXiv:2412.04862.

LG AI Research, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Yongmin Park, Sihoon Yang, Heuiyeen Yeen, Sihyuk Yi, and Hyeongu Yun. 2025b. Exaone deep: Reasoning enhanced language models. *Preprint*, arXiv:2503.12524.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2024. To the cutoff... and beyond? a longitudinal perspective on LLM

data contamination. In *The Twelfth International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024a. Kmmlu: Measuring massive multitask language understanding in korean. *Preprint*, arXiv:2402.11548.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024b. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud,

Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Google Team. 2025a. Gemma 3 technical report.

Kanana LLM Team, Yunju Bak, Hojin Lee, Minho Ryu, Jiyeon Ham, Seungjae Jung, Daniel Wontae Nam, Taegyeong Eo, Donghun Lee, Doohae Jung, Boseop Kim, Nayeon Kim, Jaesun Park, Hyunho Kim, Hyunwoong Ko, Changmin Lee, Kyoung-Woon On, Seulye Baeg, Junrae Cho, Sunghee Jung, Jieun Kang, EungGyun Kim, Eunhwa Kim, Byeongil Ko, Daniel Lee, Minchul Lee, Miok Lee, Shinbok Lee, and Gaeun Seo. 2025a. Kanana: Compute-efficient bilingual language models. *Preprint*, arXiv:2502.18934.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixing Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, Dehua Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tianshun Xing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jingyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. 2025b. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Preprint*, arXiv:2502.14739.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. Do large language model benchmarks test reliability? *Preprint*, arXiv:2502.03461.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. *Preprint*, arXiv:2504.18428.

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. 2024. 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*.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.

xAI. 2025. Grok 3 beta — the age of reasoning agents.

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. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, Donghyun Kwak, Hanock Kwak, Se Jung Kwon, Bado Lee, Dongsoo Lee, Gichang Lee, Jooho Lee, Baeseong Park, Seongjin Shin, Joonsang Yu, Seolki Baek, Sumin Byeon, Eungsup Cho, Dooseok Choe, Jeesung Han, Youngkyun Jin, Hyein Jun, Jaeseung Jung, Chanwoong Kim, Jinhong Kim, Jinuk Kim, Dokyeong Lee, Dongwook Park, Jeong Min Sohn, Sujung Han, Jiae Heo, Sungju Hong, Mina Jeon, Hyunhoon Jung, Jungeun Jung, Wangkyo Jung, Chungjoon Kim, Hyeri Kim, Jonghyun Kim, Min Young Kim, Soeun Lee, Joonhee Park, Jieun Shin, Sojin Yang, Jungsoon Yoon, Hwaran Lee, Sanghwan Bae, Jeehwan Cha, Karl Gylleus, Donghoon Ham, Mihak Hong, Youngki Hong, Yunki Hong, Dahyun Jang, Hyojun Jeon, Yujin Jeon, Yeji Jeong, Myunggeun Ji, Yeguk Jin, Chansong Jo, Shinyoung Joo, Seunghwan Jung, Adrian Jungmyung Kim, Byoung Hoon Kim, Hyomin Kim, Jungwhan Kim, Minkyoung Kim, Minseung Kim, Sungdong Kim, Yonghee Kim, Youngjun Kim, Youngkwan Kim, Donghyeon Ko, Dughyun Lee, Ha Young Lee, Jaehong Lee, Jieun Lee, Jonghyun Lee, Jongjin Lee, Min Young Lee, Yehbin Lee, Taehong Min, Yuri Min, Kiyoon Moon, Hyangnam Oh, Jaesun Park, Kyuyon Park, Younghun Park, Hanbae Seo, Seunghyun Seo, Mihyun Sim, Gyubin Son, Matt Yeo, Kyung Hoon Yeom, Wonjoon Yoo, Myungin You, Doheon Ahn, Homin Ahn, Joohee Ahn, Seongmin Ahn, Chanwoo An, Hyeryun An, Junho An, Sang-Min An, Boram Byun, Eunbin Byun, Jongho Cha, Minji Chang, Seunggyu Chang, Haesong Cho, Youngdo Cho, Dalnim Choi, Daseul Choi, Hyoseok Choi, Minseong Choi, Sangho Choi, Seongjae Choi, Wooyong Choi, Sewhan Chun, Dong Young Go, Chiheon Ham, Danbi Han, Jaemin Han, Moonyoung Hong, Sung Bum Hong, Dong-Hyun Hwang, Seongchan Hwang, Jinbae Im, Hyuk Jin Jang, Jaehyung Jang, Jaeni Jang, Sihyeon Jang, Sungwon Jang, Joonha Jeon, Daun Jeong, Joonhyun Jeong, Kyeongseok Jeong, Mini Jeong, Sol Jin, Hanbyeol Jo, Hanju Jo, Minjung Jo, Chaeyoon Jung, Hyungsik Jung, Jaeuk Jung, Ju Hwan Jung, Kwangsun Jung, Seungjae Jung, Soonwon Ka, Donghan Kang, Soyoung Kang, Taeho Kil, Areum Kim, Beomyoung Kim, Byeongwook Kim, Daehee Kim, Dong-Gyun Kim, Donggook Kim, Donghyun Kim, Euna Kim, Eunchul Kim, Geewook Kim, Gyu Ri Kim, Hanbyul Kim, Heesu Kim, Isaac Kim, Jeonghoon Kim, Jihye Kim, Joonghoon Kim, Minjae Kim, Minsub Kim, Pil Hwan Kim, Sammy Kim, Seokhun Kim, Seonghyeon Kim, Soojin Kim, Soong Kim, Soyoon Kim, Sunyoung Kim, Taeho Kim, Wonho Kim, Yoonsik Kim, You Jin Kim, Yuri Kim, Beomseok Kwon, Ohsung Kwon, Yoo-Hwan Kwon, Anna Lee, Byungwook Lee, Changho Lee, Daun Lee, Dongjae Lee, Ha-Ram Lee, Hodong Lee, Hwiyeong Lee, Hyunmi Lee, Injae Lee, Jaeung Lee, Jeongsang Lee, Jisoo Lee, Jongsoo Lee, Joongjae Lee, Juhan Lee, Jung Hyun Lee, Junghoon Lee, Junwoo Lee, Se Yun Lee, Sujin Lee, Sungjae Lee, Sungwoo Lee, Wonjae Lee, Zoo Hyun Lee, Jong Kun Lim, Kun Lim, Taemin Lim, Nuri Na, Jeongyeon Nam, Kyeong-Min Nam, Yeonseog Noh, Biro Oh, Jung-Sik Oh, Solgil Oh, Yeontaek Oh, Boyoun Park, Cheonbok Park, Dongju Park, Hyeonjin Park, Hyun Tae Park, Hyunjung Park, Jihye Park, Jooseok Park, Junghwan Park, Jungsoo Park, Miru Park, Sang Hee Park, Seunghyun Park, Soyoung Park, Taerim Park, Wonkyeong Park, Hyunjoon Ryu, Jeonghun Ryu, Nahyeon Ryu, Soonshin Seo, Suk Min Seo, Yoonjeong Shim, Kyuyong Shin, Wonkwang Shin, Hyun Sim, Woongseob Sim, Hyejin Soh, Bokyong Son, Hyunjun Son, Seulah Son, Chi-Yun Song, Chiyoung Song, Ka Yeon Song, Minchul Song, Seungmin Song, Jisung Wang, Yonggoo Yeo, Myeong Yeon Yi, Moon Bin Yim, Taehwan Yoo, Youngjoon Yoo, Sungmin Yoon, Young Jin Yoon, Hangyeol Yu, Ui Seon Yu, Xingdong Zuo, Jeongin Bae, Joungeun Bae, Hyunsoo Cho, Seonghyun Cho, Yongjin Cho, Taekyoon Choi, Yera Choi, Jiwan Chung, Zhenghui Han, Byeongho Heo, Euisuk Hong, Taebaek Hwang, Seonyeol Im, Sumin Jegal, Sumin Jeon, Yelim Jeong, Yonghyun Jeong, Can Jiang, Juyong Jiang, Jiho Jin, Ara Jo, Younghyun Jo, Hoyoun Jung, Juyoung Jung, Seunghyeong Kang, Dae Hee Kim, Ginam Kim, Hangyeol Kim, Heeseung Kim, Hyojin Kim, Hyojun Kim, Hyun-Ah Kim, Jeehye Kim, Jin-Hwa Kim, Jiseon Kim, Jonghak Kim, Jung Yoon Kim, Rak Yeong Kim, Seongjin Kim, Seoyoon Kim, Sewon Kim, Sooyoung Kim, Suky-

oung Kim, Taeyong Kim, Naeun Ko, Bonseung Koo, Heeyoung Kwak, Haena Kwon, Youngjin Kwon, Boram Lee, Bruce W. Lee, Dagyeong Lee, Erin Lee, Euijin Lee, Ha Gyeong Lee, Hyojin Lee, Hyunjeong Lee, Jeeyoon Lee, Jeonghyun Lee, Jongheok Lee, Joonhyung Lee, Junhyuk Lee, Mingu Lee, Nayeon Lee, Sangkyu Lee, Se Young Lee, Seulgi Lee, Seung Jin Lee, Suhyeon Lee, Yeonjae Lee, Yesol Lee, Youngbeom Lee, Yujin Lee, Shaodong Li, Tianyu Liu, Seong-Eun Moon, Taehong Moon, Max-Lasse Nihlenramstroem, Wonseok Oh, Yuri Oh, Hongbeen Park, Hyekyung Park, Jaeho Park, Nohil Park, Sangjin Park, Jiwon Ryu, Miru Ryu, Simo Ryu, Ahreum Seo, Hee Seo, Kangdeok Seo, Jamin Shin, Seungyoun Shin, Heetae Sin, Jiangping Wang, Lei Wang, Ning Xiang, Longxiang Xiao, Jing Xu, Seonyeong Yi, Haanju Yoo, Haneul Yoo, Hwanhee Yoo, Liang Yu, Youngjae Yu, Weijie Yuan, Bo Zeng, Qian Zhou, Kyunghyun Cho, Jung-Woo Ha, Joonsuk Park, Jihyun Hwang, Hyoung Jo Kwon, Soonyong Kwon, Jungyeon Lee, Seungho Lee, Seonghyeon Lim, Hyunkyung Noh, Seungho Choi, Sang-Woo Lee, Jung Hwa Lim, and Nako Sung. 2024. Hyperclova x technical report. *Preprint*, arXiv:2404.01954.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *Preprint*, arXiv:2411.09116.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. 2024. Mmlu-cf: A contamination-free multi-task language understanding benchmark. *Preprint*, arXiv:2412.15194.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024. SGLang: Efficient execution of structured language model programs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Dataset Examples

The Table 4 presents the example of KMMLU-REDUXand KMMLU-PRO. While both benchmarks include domains such as management and accounting, KMMLU-REDUX emphasizes industrial expert knowledge, whereas KMMLU-PRO focuses on practical professional expertise. This allows for a clear comparison of the different knowledge evaluations between the two benchmarks.

B Details of KMMLU-REDUX Errors

B.1 Error Statistics

We define a set of error types based on recurring issues observed during our analysis of the KMMLU. Then, we find out the number of data instances per each error type as shown in Table 5. To identify leaked answer cases, we apply rule-based filtering using string-overlap heuristics. For other error types, including notation errors, bad clarity, and ill-posed questions, we leverage GPT-4o to assist in annotation. Also, we provide the prompt used for annotation in Figure 6.

- **Ill-posed Question** : The question lacks critical references or contextual information.
- **Leaked Answer** : The ground truth is explicitly stated within the question itself.
- **Notation Error** : Errors in mathematical expressions or chemical equations.
- **Bad Clarity** : The data itself is unclear and contains grammatical errors.

B.2 Example of Error Types

Table 12 provides examples of the error types in 2.2 identified in KMMLU. Each example illustrates a specific issue that affects benchmark reliability.

C Korean National Technical Qualification List of KMMLU-REDUX

Table 11 presents the list of Korean National Technical Qualifications (KNTQs) included in our benchmark along with their corresponding official exam dates. To facilitate analysis, we categorize the 100 NTQs into Korean Standard Industrial Classification (KSIC) where mapping to the exam (see Figure 7). This categorization enables structured evaluation across diverse domains and better reflects the real-world industrial fields.

	KMMLU-REDUX	KMMLU-PRO
Question	전수검사가 불가능하여 반드시 샘플링검사를 하여야 하는 경우는? *Which case is necessary to conduct sampling inspections due to a complete inspection is not possible?*	액면주식의 주권을 발행한 비상장주식회사의 상법 제329조의2 소정의 주식분할에 관한 설명으로 옳은 것은? *Which of the following statements about stock splits under Article 329-2 of the Commercial Act for unlisted stock companies that have issued share certificates is correct?*
Options	["전기제품의 출력전압의 측정", "주물제품의 내경가공에서 내경의 측정", "전구의 수입검사에서 전구의 점등시험", "진공관의 수입검사에서 진공관의 평균수명 추정"] *["Measuring the output voltage of electrical products", "Measuring the inner diameter in machining of cast products", "Lighting test for imported light bulbs during inspection", "Estimating the average lifespan of vacuum tubes during import inspection"]*	["주식분할을 하기 위해서는 주주총회의 특별결의를 거쳐야 한다.", "회사가 공고한 주권제출기간 중에 주주가 주권을 제출하면 그 시점에 주식분할의 효력이 발생한다.", "주식분할이 이루어져도 발행주식총수는 증가하지 않는다.", "주식분할이 이루어져도 1주의 액면금액은 감소하지 않는다.", "주식발행이 이루어지면 회사의 자본금이 증가한다."] *["A special resolution at the shareholders' meeting is required to carry out a stock split.", "If a shareholder submits share certificates during the public notice period announced by the company, the stock split becomes effective at that time.", "The total number of issued shares does not increase even after a stock split.", "The face value per share does not decrease after a stock split.", "When shares are issued, the company's capital increases."]*
Answer	진공관의 수입검사에서 진공관의 평균수명 추정 *Estimating the average lifespan of vacuum tubes during import inspection*	주식분할을 하기 위해서는 주주총회의 특별결의를 거쳐야 한다. *A special resolution at the shareholders' meeting is required to carry out a stock split.*
License name	품질경영기사 *Engineer Quality Management*	회계사 *Certified Public Accountant*

Table 4: Examples of KMMLU-Redux and KMMLU-Pro. Gray text represents English translations of the original Korean.

Error Type	# of Questions (Ratio)
Ill-posed Question	512 (1.46 %)
Leaked Answer	42 (0.11%)
Notation Error	846 (2.42%)
Bad Clarity	1284 (3.67%)

Table 5: Statistics of the error types

D Details of KMMLU-PRO Annotation

D.1 Annotation Pipeline

We first conduct OCR parsing with GPT-4o on PDF files of KNPL acquisition exams. With the parsed data, the main tasks for human annotators are: 1) reviewing parsing errors, 2) converting tables into latex format, and 3) converting images into text which conveys same meaning. If it is impossible to convert an image into text, we remove it. For the cases where multiple answers are allowed, commonly due to the ambiguity of question itself, we discard them.

As we leverage the official PDF files managed and controlled by the government, we can guarantee the correctness of answer label. This help us

GPT-4o Error Annotation Prompt

You are a helpful assistant that annotates error types in questions.
- Ill-posed question stands for which question miss the critical reference information to solve the question (such as table, formular, image, etc.).
- Notation correct stands for which question has correct math and chemical notation. Criteria of correctness whether the notation is impact to solve the question. (e.g. m2 is incorrect, but $m\hat{2}$ is correct. 센티미터 is correct.)
- Grammar correct stands for which question has correct grammar and spelling. Criteria of correctness whether the notation is impact to understand the question. (e.g. 상 담 기 법 보 다 는 상 담 자 의 인 간 적 자 질 과 진 솔 한 태 도 를 중 시 한 다 . is grammatically incorrect due to spacing error.)

Please check the following question whether it has error types.
Check ill-posed question True/False.
Check notation correct True/False.
Check grammar correct True/False.
Then, if there is any error, please explain the reason.

Question
{{*question*}}

Figure 6: The prompt is used for error type annotation. Each sample is annotated as an error if the respective field returns True. The 'Grammar Correct' field is used to detect 'Bad Clarity' cases.

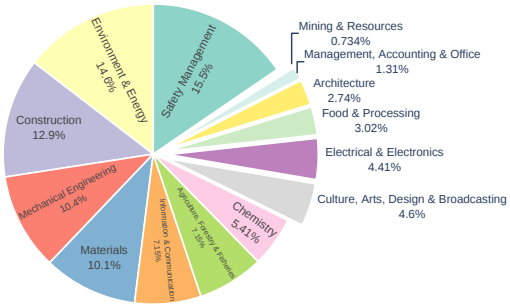

Figure 7: Domain distribution of problems in KMMLU-REDUX. The total size of the dataset is 2,587.

to save costs because we do not need experts for annotations nor the answer relabeling to avoid risk of data from online (Gema et al., 2025; Team et al., 2025b). Before annotation, we explained the context of benchmark construction about the Korean professional license exams to human annotators. We present annotation instructions in Fig 8.

D.2 Annotators Demographics

Category	Demographics	Counts
Gender	Female	23
	Male	-
Age	20s	13
	30s	8
	40s	2
Academic background	Bachelor's degrees	20
	Master's degrees	3

Table 6: Demographics of the human annotators for KMMLU-PRO.

With 23 annotators, it took 8 business days to complete all annotations, including project setup. The detailed demographics of annotators are pre-

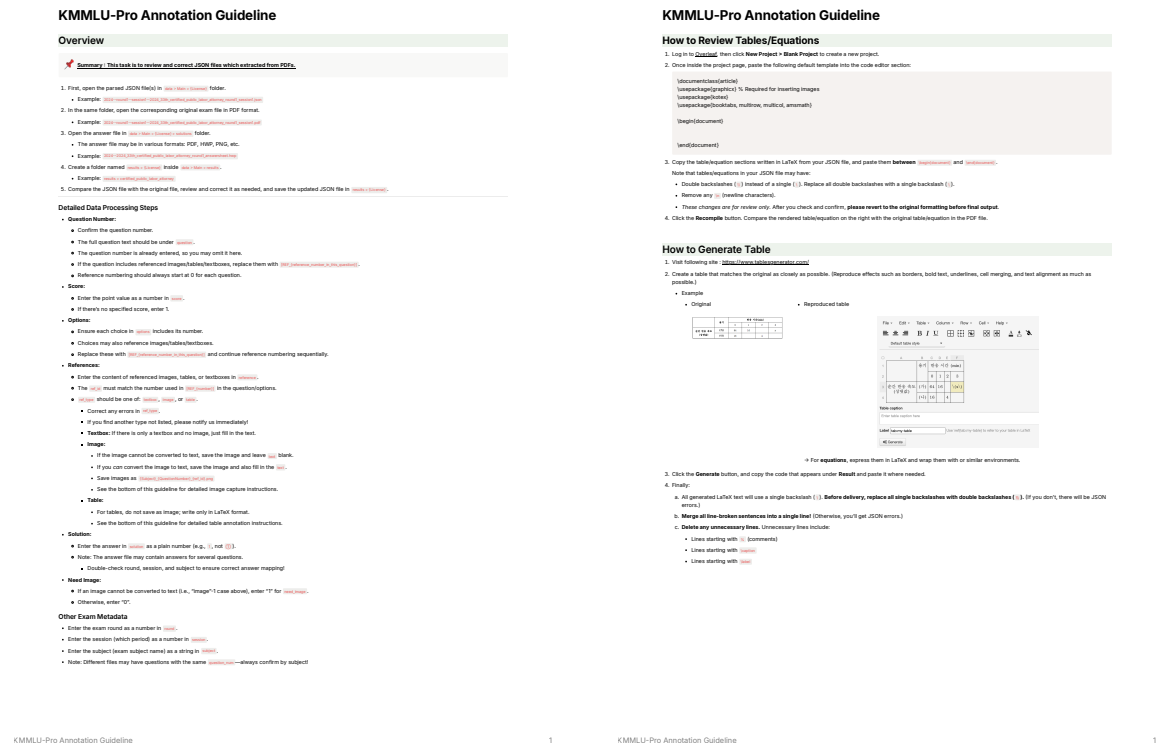

Figure 8: Excerpt from the translated annotation guidelines for converting PDF documents into structured text. We carefully instruct LaTeX table formatting.

sented in Table 6. The total amount of annotations was approximately $14,000. The average hourly wage is 8.83 U.S. dollars, which is higher than the legal minimum wage at the time of hiring in South Korea.

E Evaluation Setup

E.1 Evaluation Prompts

The figure 9 and figure 10 present the prompt for the evaluation written in English and Korean, respectively. For the English prompt, we use the regex expression of `r"(?i)Answer[^A-E]*:[^A-E]*([A-E])"`. For the Korean prompt, we use `r"정답[^A-E]*:[^A-E]*([A-E])"`. The regex expressions for the flexible parsing are `r"Answer[^A-E]*([A-E])|([A-E])\)"` and `r"정답[^A-E]*([A-E])|([A-E])\)"`, respectively.

E.2 Inference Engines

Excluding closed models, the main inference engine we use is SGLang (Zheng et al., 2024). However, for the Gemma 3 series and Mistral Small 3.1 Instruct models, we adopt vLLM (Kwon et al., 2023) due to their incompatibility with SGLang at the time of the paper writing.

E.3 License Passing Criteria for KMMLU-Pro

We follow the official scoring criteria used for each license examination. All licensing exams in KMMLU-Pro are composed of multiple subjects. Candidates are typically required to score at least 40% in every subject and achieve an average score of at least 60%, except for the cases of the Certified Judicial Scrivener and the Lawyer. For the Certified Judicial Scrivener exam, candidates need only score at least 40% in each subject, with no requirement regarding the average. The Lawyer license exam uses a relative grading system where only a certain proportion of top-scoring candidates pass. The usual cut-off point is approximately 54.22 (900 out of 1660), which we use as the passing threshold.

It is also important to note that our evaluation benchmark is text-based, not multi-modal. Therefore, we exclude questions that include images. In addition, candidates often need to go through multiple exam stages to obtain a license, with the later stages, such as the second or third, containing de-

English Prompt

Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering.

{{*question*}}

A) {{*option A*}}
B) {{*option B*}}
C) {{*option C*}}
D) {{*option D*}}

Figure 9: The English prompt used for evaluating LLMs on our KMMLU-REDUX and KMMLU-PRO. This prompt is exactly same with the prompt used for Multiple-Choices Question Answering (MCQA) in OpenAI's simple-evals repository. The number of options is adjusted for each problems.

Korean Prompt

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

{{*question*}}

A) {{*option A*}}
B) {{*option B*}}
C) {{*option C*}}
D) {{*option D*}}

Figure 10: The Korean prompt used for evaluating LLMs on our KMMLU-REDUX and KMMLU-PRO. This prompt is translated version of the English prompt. The number of options is adjusted for each problems.

scriptive questions. However, since we only collect multiple-choice questions, the descriptive problems are excluded. Lastly, we exclude questions with multiple answers introduced by the ambiguity of the question.

F Detailed Results

F.1 The Results for Smaller (<10B) Models

The Table 7 presents the results of KMMLU-PRO and KMMLU-REDUX for smaller (<10B) models. Since many of the *tiny* models in this table were used to construct KMMLU-REDUX through adversarial filtration, their KMMLU-REDUX scores are biased. Nevertheless, as shown by the results for larger models, models equipped with dense reasoning capabilities usually outperform their counterparts without reasoning (e.g., Qwen3-8B with and without "thinking").

F.2 Breakdown of Results of KMMLU-REDUX

The Table 8 presents the breakdown results for 14 categories in KMMLU-REDUX across various LLMs.

F.3 Breakdown of Results of KMMLU-PRO

The Table 9 shows the breakdown results for all NPLs in KMMLU-PRO across various LLMs. While the models relatively easily pass licensing in the medicine domain, they struggle in the Law and Tax&Accounting domains.

	KMMLU-Redux	KMMLU-Pro		
	Acc	Acc	# of passed KNPLs	Avg. Acc (micro)
DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025a)	21.30*	20.55	0/14	20.91
Llama 3.2 3B Instruct (Meta, 2024c)	17.59*	25.53	0/14	21.73
Gemma 3 4B IT (Team, 2025a)	25.09*	32.86	0/14	29.14
Qwen 2.5 3B Instruct (Qwen et al., 2025)	24.74*	33.27	0/14	29.19
Qwen3-1.7B (Yang et al., 2025)	28.99	30.42	0/14	29.74
Kanana Nano 2.1B Instruct (Team et al., 2025a)	27.25*	32.60	0/14	30.04
Aya Expanse 8B (Dang et al., 2024)	28.30	31.65	0/14	30.05
HyperCLOVAX-SEED-Text-Instruct-1.5B (naver hyperclovax, 2025)	33.94	30.13	0/14	31.95
Llama 3.1 8B Instruct (Meta, 2024b)	31.89	33.81	0/14	32.89
Qwen3-1.7B (w/ thinking) (Yang et al., 2025)	37.80	38.27	1/14	38.05
EXAONE 4.0 1.2B (Research et al., 2025a)	40.43	37.48	0/14	38.89
Ko-R1-7B-v2.1 (OneLineAI, 2025)	41.94	38.70	1/14	40.25
EXAONE 4.0 1.2B (w/ thinking) (Research et al., 2025a)	46.85	42.69	0/14	44.68
Qwen3-8B (Yang et al., 2025)	49.25	46.92	1/14	48.03
Qwen3-8B (w/ thinking) (Yang et al., 2025)	58.79	55.27	3/14	56.95

Table 7: The main evaluation results of KMMLU-REDUX and KMMLU-PRO benchmarks on smaller (< 10B) LLMs. The gray-shaded models are the dense-reasoning models. The KMMLU-REDUX scores with * are biased as these models are used for the dataset filtration (Section 2.2.1).

F.4 Performance Comparison Between Reasoning and Non-Reasoning Models

To further analyze the effects of *reasoning* (or *thinking*; we use these two terms interchangeably), we compare the scores of reasoning models with their corresponding non-reasoning models for each license in KMMLU-PRO. The Qwen3 series, EXAONE 4.0 series, and Claude 3.7 support both think-on/off modes, so we compare the scores under both modes. Other closed models, such as OpenAI's GPT-4.1 vs. O-series, are excluded since we do not have information whether the reasoning models share the same architecture with instruction-tuned models. Most of the scores are from Table 9, except for DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI et al., 2025a), which we evaluated newly. Table 10 presents the performance gains (%) when reasoning is enabled.

Domain		*Safety Management*	*Environment & Energy*	*Construction*	*Mechanical Engineering*	*Materials*	*Information & Communication*	*Agriculture, Forestry & Fisheries*	*Chemistry*	*Culture, Arts, Design & Broadcasting*	*Electrical & Electronics*	*Food & Processing*	*Architecture*	*Management, Accounting & Office*	*Mining & Resources*	*Avg.*
Open-weight Models																
< 5B	DeepSeek-R1-Distill-Qwen-1.5B	20.50	17.77	24.32	20.00	24.43	26.49	17.84	20.71	19.33	20.18	15.38	18.31	20.59	31.58	21.30*
	Llama 3.2 3B Instruct	13.00	12.20	13.81	14.81	11.07	15.68	16.76	11.43	17.65	7.89	16.67	11.27	5.88	15.79	17.59*
	Qwen 2.5 3B Instruct	16.50	17.24	15.32	17.41	15.27	20.54	18.38	11.43	26.05	16.67	11.54	15.49	14.71	10.53	24.74*
	Gemma 3 4B IT	23.00	23.87	22.52	25.56	20.23	28.65	32.43	25.71	21.01	27.19	21.79	21.13	14.71	10.53	25.09*
	Kanana Nano 2.1B Instruct	24.75	27.32	25.83	26.67	29.77	22.16	26.49	23.57	25.21	28.07	28.21	28.17	26.47	10.53	27.25*
	Qwen3-1.7B	30.50	27.06	27.93	36.67	32.82	29.73	23.24	30.00	23.53	28.95	25.64	23.94	20.59	15.79	28.99
	HyperCLOVAX-SEED-Text-Instruct-1.5B	28.00	25.99	30.93	32.59	33.21	29.19	32.97	22.86	35.29	28.07	32.05	32.39	17.65	26.32	33.94
	Qwen3-1.7B (w/ thinking)	37.25	35.54	33.03	39.63	40.08	41.62	36.76	40.71	35.29	46.49	42.31	29.58	41.18	42.11	37.80
	EXAONE 4.0 1.2B	44.00	37.14	39.34	44.07	44.27	48.11	38.38	34.29	41.18	35.96	32.05	38.03	29.41	21.05	40.43
	EXAONE 4.0 1.2B (w/ thinking)	47.50	44.03	49.30	47.41	50.76	50.81	42.70	50.00	44.54	50.00	38.46	49.30	32.35	21.05	46.85
< 10B	Aya Expanse 8B	25.00	20.42	23.72	21.11	24.05	29.19	29.19	21.43	23.53	22.81	20.51	30.99	26.47	21.05	28.30
	Llama 3.1 8B Instruct	30.50	25.99	23.12	28.15	30.53	33.51	32.97	22.14	35.29	24.56	29.49	26.76	29.41	31.58	31.89
	Ko-R1-7B-v2.10	42.50	37.93	39.94	41.48	38.55	50.81	37.30	52.14	49.58	50.00	33.33	33.80	50.00	36.84	41.94
	Qwen3-8B	47.00	48.01	43.84	49.63	56.11	56.22	41.62	58.57	47.06	60.53	46.15	43.66	50.00	31.58	49.25
	Qwen3-8B (w/ thinking)	55.75	57.82	53.15	67.78	64.89	62.16	45.95	65.71	57.14	66.67	56.41	53.52	58.82	63.16	58.79
< 20B	Gemma 3 12B IT	44.50	42.18	46.25	44.81	46.95	50.81	51.35	50.71	57.98	48.25	46.15	39.44	50.00	42.11	46.70
	Phi-4 (14B)	48.50	47.75	48.05	52.59	49.62	52.43	44.86	58.57	54.62	56.14	50.00	36.62	52.94	36.84	49.75
	Qwen3-14B	52.75	55.70	51.95	63.33	61.45	65.95	47.57	61.43	57.98	67.54	52.56	59.15	61.76	47.37	57.25
	Qwen3-14B (w/ thinking)	64.50	66.58	60.36	70.00	72.90	68.65	54.05	73.57	63.87	75.44	51.28	57.75	70.59	68.42	65.71
< 32B	Aya Expanse 32B	30.50	32.10	27.03	38.15	35.50	39.46	37.84	28.57	33.61	35.09	32.05	35.21	26.47	21.05	33.05
	Mistral Small 3.1 Instruct (24B)	45.00	50.40	51.95	55.93	59.54	58.38	51.35	58.57	56.30	58.77	47.44	52.11	58.82	31.58	52.92
	Gemma 3 27B IT	49.50	51.19	47.45	57.41	58.02	62.70	49.73	60.71	59.66	62.28	56.41	47.89	67.65	31.58	54.04
	Qwen3-30B-A3B	54.25	57.56	54.65	58.52	64.50	61.08	51.89	65.00	62.18	68.42	58.97	56.34	58.82	52.63	58.41
	EXAONE 4.0 32B	61.00	62.86	61.26	70.74	74.05	70.81	60.54	67.86	74.79	63.16	55.13	49.30	61.76	42.11	64.79
	Qwen3-32B	59.50	64.19	60.66	68.52	72.52	69.19	58.38	68.57	74.79	72.81	58.97	53.52	70.59	63.16	64.98
	Qwen3-30B-A3B (w/ thinking)	63.00	64.19	62.46	69.63	70.23	69.19	56.76	69.29	68.91	72.81	53.85	56.34	70.59	68.42	65.25
	QwQ 32B	61.25	67.64	68.17	71.11	74.05	72.97	58.38	71.43	72.27	71.05	53.85	56.34	79.41	52.63	67.34
	Qwen3-32B (w/ thinking)	64.50	68.97	66.97	74.07	75.57	70.81	62.70	74.29	68.91	72.81	64.10	53.52	73.53	57.89	68.77
	EXAONE 4.0 32B (w/ thinking)	70.50	71.35	68.17	74.44	83.21	76.76	68.65	74.29	77.31	74.44	69.23	57.75	82.35	52.63	72.71
> 70B	Llama 3.3 70B Instruct	52.25	54.11	52.85	62.59	59.92	64.32	52.43	50.71	60.50	60.53	55.13	53.52	64.71	36.84	56.17
	C4AI Command A (111B)	56.75	66.58	63.06	63.33	64.89	67.57	62.16	60.71	68.91	64.04	62.82	59.15	55.88	47.37	62.93
	DeepSeek V3 (671B)	62.50	62.33	63.66	66.30	72.52	72.97	62.70	66.43	67.23	69.30	69.23	59.15	61.76	63.16	65.64
	Llama-4-Scout-17B-16E-Instruct	64.00	67.64	65.77	69.63	77.10	71.35	61.08	69.29	66.39	71.93	62.82	57.75	64.71	57.89	67.49
	Qwen3-235B-A22B	64.25	72.68	66.37	71.11	82.06	72.43	65.41	71.43	68.07	73.68	65.38	52.11	67.65	47.37	69.54
	Qwen3-235B-A22B (w/ thinking)	69.75	75.86	71.47	81.11	85.11	76.22	64.86	75.00	70.59	78.95	73.08	61.97	82.35	68.42	74.49
	Llama-4-Maverick-17B-128E-Instruct	73.50	76.13	78.38	79.26	83.97	80.54	77.30	75.00	78.15	79.82	73.08	71.83	82.35	73.68	77.58
	DeepSeek R1 (671B)	72.50	76.39	79.58	80.74	85.11	81.62	82.70	77.14	79.83	81.58	75.64	63.38	85.29	73.68	78.51
Closed Models																
	GPT-4.1 mini (2024-04-14)	61.00	64.99	63.66	69.26	74.81	72.43	63.78	73.57	70.59	66.67	75.64	56.34	73.53	57.89	67.03
	o3-mini (2025-01-31)	63.75	66.05	63.96	72.96	73.66	75.14	63.78	69.29	72.27	75.44	66.67	46.48	76.47	57.89	67.84
	Grok-3-mini-beta	69.75	69.76	69.07	73.33	83.97	77.30	65.95	70.71	68.07	73.68	61.54	59.15	76.47	73.68	71.47
	Grok-3-beta	70.00	70.56	68.47	76.30	83.97	76.76	69.19	75.71	74.79	69.30	70.51	71.83	73.53	57.89	72.90
	o4-mini (2025-04-16)	67.75	73.74	76.28	78.52	82.06	81.08	76.76	78.57	80.67	79.82	71.79	61.97	79.41	78.95	75.80
	GPT-4.1	71.50	74.27	72.07	77.41	85.11	81.62	80.00	74.65	76.47	77.19	75.64	61.97	76.47	68.42	75.86
	Claude 3.7 Sonnet	70.75	76.92	76.28	80.37	83.97	77.84	78.92	78.57	76.47	78.07	75.64	64.79	79.41	68.42	76.88
	Claude 3.7 Sonnet (w/ thinking)	71.50	80.37	79.88	80.74	87.02	82.16	80.00	82.86	80.67	79.82	73.08	70.42	85.29	68.42	79.36
	o3	77.75	78.25	77.78	80.37	85.50	78.92	83.24	78.17	84.03	85.09	82.05	66.20	85.29	78.95	79.92
	o1	75.75	79.58	81.98	81.11	90.84	80.54	84.86	77.86	84.03	81.58	83.33	70.42	85.29	73.68	81.14

Table 8: The breakdown results for 14 categories in KMMLU-REDUX. The gray-shaded models are the dense-reasoning models. The scores with * are biased as these models are used for the dataset filtration (Section 2.2.1).

Domain		Law				Tax & Accounting			Value Estimation		Medical						
	Names of KNPLs	*Judicial Scrivener*	*Lawyer*	*Public Labor Attorney*	*Patent Attorney*	*Public Accountant (CPA)*	*Tax Accountant*	*Customs Broker*	*Damage Adjuster (CDA)*	*Appraiser*	*Doctor of Korean Medicine*	*Dentist*	*Pharmacist*	*Herb Pharmacist*	*Physician*	*Avg.*	*# of passed NPLs*
Open-weight Models																	
< 5B	DeepSeek-R1-Distill-Qwen-1.5B	15.66	13.33	24.27	25.69	23.56	18.07	18.24	20.00	21.94	20.49	18.49	23.25	23.77	18.67	20.55	0/14
	Llama 3.2 3B Instruct	27.27	24.67	25.94	20.18	16.35	23.95	23.9	23.33	19.39	23.26	26.45	36.53	30.74	28.67	25.53	0/14
	HyperCLOVAX-SEED-Text-Instruct-1.5B	24.24	24.67	30.13	26.61	25.48	29.83	30.19	41.67	25.00	30.90	27.31	36.16	35.66	33.33	30.13	0/14
	Qwen3-1.7B	19.19	20.67	30.96	32.11	21.63	21.01	23.9	30.0	29.08	30.56	30.32	46.49	46.72	33.33	30.42	0/14
	Kanana Nano 2.1B Instruct	19.19	24.67	30.13	22.02	23.08	26.47	26.42	28.33	26.53	42.01	31.61	52.40	44.67	38.67	32.60	0/14
	Gemma 3 4B IT	23.23	26.67	32.22	26.61	27.4	26.47	27.67	29.17	23.47	37.15	36.56	49.08	43.03	36.00	32.86	0/14
	Qwen 2.5 3B Instruct	20.20	25.33	35.15	24.77	21.63	23.11	30.19	35.00	27.04	37.85	35.05	47.97	52.46	34.67	33.27	0/14
	EXAONE 4.0 1.2B (w/ thinking)	26.26	22.00	35.98	25.69	30.77	41.60	32.70	44.17	37.24	41.80	37.63	54.24	41.80	44.67	37.48	0/14
	Qwen3-1.7B (w/ thinking)	25.25	25.33	31.8	34.86	25.48	25.63	29.56	34.17	36.22	44.79	38.06	**61.99**	59.43	44.67	38.27	1/14
	EXAONE 4.0 1.2B (w/ thinking)	28.29	26.00	39.75	37.61	38.94	36.55	43.40	53.33	43.88	42.71	39.35	63.47	47.13	50.67	42.69	0/14
< 10B	Aya Expanse 8B	20.20	20.00	30.54	22.02	23.56	18.49	37.11	33.33	22.45	36.11	29.25	50.18	46.31	42.00	31.65	0/14
	Llama 3.1 8B Instruct	28.28	21.33	30.96	23.85	21.15	23.53	33.33	33.33	26.02	35.42	33.76	54.61	50.82	42.00	33.81	0/14
	Ko-R1-7B-v2.1	40.59	28.85	38.33	39.8	35.22	33.94	26.47	29.86	42.21	29.29	48.67	46.24	**64.21**	30.67	38.70	1/14
	Qwen3-8B	29.29	34.0	46.03	35.78	33.17	35.71	34.59	43.33	40.82	50.69	46.02	71.96	**73.36**	59.33	46.92	1/14
	Qwen3-8B (w/ thinking)	27.78	36.0	49.37	41.28	48.08	44.12	49.69	50.83	53.06	61.81	54.19	**83.03**	**76.23**	**75.33**	55.27	3/14
< 20B	Phi-4 (14B)	33.33	34.00	41.00	36.70	37.50	37.82	42.77	44.17	43.37	45.49	41.29	**72.69**	57.38	51.33	45.32	1/14
	Gemma 3 12B IT	28.79	23.33	40.59	39.45	34.62	34.87	42.14	39.17	35.71	52.08	49.46	71.59	**62.30**	**68.00**	45.82	2/14
	Qwen3-14B	32.83	30.67	39.75	47.71	48.56	38.24	42.14	49.17	50.00	57.64	55.91	**81.18**	**75.0**	**75.33**	53.02	3/14
	Qwen3-14B (w/ thinking)	30.81	36.67	48.95	47.71	58.65	51.26	52.2	47.50	54.59	65.28	61.94	**87.82**	**80.74**	**82.67**	59.48	3/14
< 32B	Aya Expanse 32B	24.75	18.67	24.27	32.11	20.67	21.01	27.67	34.17	23.98	32.64	31.83	53.14	45.49	38.67	31.26	0/14
	Mistral Small 3.1 Instruct (24B)	27.27	33.33	43.51	39.45	39.90	38.66	42.77	47.50	41.33	52.08	49.89	**79.34**	**68.03**	**72.00**	49.49	3/14
	Gemma 3 27B IT	29.80	31.33	44.35	33.94	38.46	43.28	40.88	43.33	44.39	52.08	58.49	81.18	**73.36**	**72.67**	51.03	2/14
	Qwen3-30B-A3B	31.31	26.00	47.28	35.78	45.19	35.71	44.03	49.17	47.96	56.60	53.98	**83.03**	**77.87**	**72.00**	52.33	3/14
	Qwen3-32B	33.33	35.33	53.14	43.12	57.69	45.80	53.46	53.33	48.98	61.46	60.86	**85.24**	**84.84**	**84.00**	58.86	3/14
	EXAONE 4.0 32B	39.90	43.33	57.32	50.46	54.81	53.36	49.69	57.50	55.61	60.07	58.49	**83.76**	**77.05**	**84.00**	60.01	3/14
	Qwen3-30B-A3B (w/ thinking)	35.35	37.33	50.21	45.87	57.21	50.0	48.43	54.17	54.08	70.49	59.78	**88.19**	**84.43**	**84.67**	60.52	3/14
	Qwen3-32B (w/ thinking)	33.33	34.67	55.23	43.12	55.29	47.06	54.09	57.50	55.61	70.14	**66.88**	87.08	**81.56**	**88.67**	61.14	3/14
	QwQ 32B	35.35	39.33	55.65	44.04	60.58	55.04	57.23	56.67	**64.29**	71.53	**66.24**	**88.93**	**84.43**	**88.67**	63.94	5/14
	EXAONE 4.0 32B (w/ thinking)	41.92	39.33	60.25	59.63	65.38	57.56	**66.67**	**64.17**	**69.90**	75.69	**68.60**	88.56	**83.61**	**87.33**	67.67	6/14
> 70B	Llama 3.3 70B Instruct	32.83	41.33	46.44	42.20	42.31	38.66	49.69	54.17	43.37	51.39	56.77	**85.98**	**71.72**	**74.00**	53.24	3/14
	C4AI Command A (111B)	41.92	34.00	51.46	49.54	45.67	45.80	46.54	54.17	52.04	61.46	57.63	**83.39**	**79.10**	**80.67**	57.48	3/14
	Llama-4-Scout-17B-16E-Instruct	35.86	34.00	53.14	50.46	47.12	43.70	47.80	54.17	47.45	61.46	**68.17**	**81.92**	**85.25**	**82.67**	58.14	4/14
	DeepSeek V3 (671B)	38.89	36.00	56.49	48.62	50.00	42.44	47.80	52.50	53.06	**71.53**	64.95	**87.82**	**87.70**	**84.67**	60.77	4/14
	Qwen3-235B-A22B	38.38	39.33	54.81	45.87	55.77	49.58	52.83	**60.0**	57.65	64.24	**70.97**	86.72	**86.07**	**84.67**	62.12	4/14
	Llama-4-Maverick-17B-128E-Instruct	38.38	41.33	62.34	55.96	61.06	56.72	56.60	**69.17**	66.84	75.00	**76.34**	89.67	**90.98**	**90.67**	68.10	4/14
	Qwen3-235B-A22B (w/ thinking)	43.43	42.67	56.90	57.80	69.23	61.76	61.64	59.17	**62.76**	**71.88**	**77.42**	**89.67**	**89.34**	**88.00**	68.22	6/14
	DeepSeek R1 (671B)	45.45	38.00	61.51	57.80	66.83	60.50	**69.18**	**68.33**	**64.80**	79.51	**81.29**	**92.99**	**94.67**	**92.67**	71.33	7/14
Closed Models																	
	GPT-4.1 mini (2024-04-14)	38.38	32.00	56.90	56.88	54.81	48.32	50.31	50.00	55.10	67.36	**72.47**	**90.77**	**81.97**	**90.00**	62.18	4/14
	o3-mini (2025-01-31)	38.38	38.67	51.46	47.71	60.10	47.06	50.31	52.50	57.65	64.24	**76.56**	88.93	**80.33**	**91.33**	62.05	3/14
	Grok-3-mini-beta	37.37	34.67	57.74	48.62	67.79	49.58	**65.41**	54.17	64.80	66.32	**73.98**	**91.14**	**84.84**	**90.0**	65.08	5/14
	Grok-3-beta	42.42	41.33	59.00	55.96	63.94	57.98	**62.89**	**61.67**	**66.84**	70.49	**77.85**	**89.30**	**91.80**	**94.67**	68.37	7/14
	o4-mini (2025-04-16)	37.37	46.0	62.34	49.54	69.23	55.04	**66.67**	59.17	**67.35**	76.74	**82.15**	**93.36**	**89.75**	**92.0**	69.65	6/14
	GPT-4.1	47.47	50.00	**66.95**	**63.30**	66.83	**59.24**	**67.30**	**66.67**	**68.37**	80.21	**82.80**	**94.10**	**92.62**	**94.67**	72.99	10/14
	o3	45.96	41.33	71.13	57.80	73.56	**61.76**	**70.44**	**65.83**	**71.94**	**77.43**	**87.96**	**92.99**	**91.39**	**94.67**	73.60	9/14
	Claude 3.7 Sonnet	55.56	53.33	**73.22**	**63.30**	68.27	**64.29**	**67.92**	**77.50**	**73.47**	70.83	**81.51**	**94.46**	**92.21**	**93.33**	74.52	10/14
	Claude 3.7 Sonnet (w/ thinking)	50.00	**56.00**	**77.41**	**74.31**	78.85	**70.17**	**75.47**	**70.00**	**81.12**	**76.39**	**84.09**	**93.36**	**93.03**	**92.67**	77.70	12/14
	o1	54.55	49.33	**71.55**	**65.14**	75.48	67.23	**76.73**	**78.33**	**78.06**	**83.33**	**88.39**	**94.10**	**95.49**	**96.67**	78.09	10/14

Table 9: The break down results for all KNPLs in KMMLU-Pro. The gray-shaded models are the dense-reasoning models. The **blue** scores indicate that the LLM obtain the license. The details for the pass criteria of each license are described in Appendix E.3.

Domain	Law				Tax & Accounting			Value Estimation		Medical				
Names of KNPLs	*Judicial Scrivener*	*Lawyer*	*Public Labor Attorney*	*Patent Attorney*	*Public Accountant (CPA)*	*Tax Accountant*	*Customs Broker*	*Damage Adjuster (CDA)*	*Appraiser*	*Doctor of Korean Medicine*	*Dentist*	*Pharmacist*	*Herb Pharmacist*	*Physician*
EXAONE 4.0 1.2B	7.73	18.18	10.48	46.40**	26.55*	-12.14	32.72**	20.74*	17.83	2.18	4.57	17.02**	12.75	13.43
Qwen3-1.7B	31.58*	22.58	2.70	8.57	17.78	22.00	23.68	13.89	24.56*	46.59**	25.53**	33.33**	46.59**	34.00**
Qwen3-8B	-5.17	5.88	7.27	15.38	44.93**	23.53**	43.64**	17.31	30.00**	21.92**	17.76**	15.38**	21.92**	26.97**
Qwen3-14B	7.69	10.87	23.16**	3.85	21.78**	25.27**	31.34**	0.00	10.20	15.66**	15.00**	7.73**	15.66**	7.96
Qwen3-30B-A3B	12.90	43.59**	6.19	28.21*	26.60**	40.00**	10.00	10.17	12.77	24.54**	10.76	6.22*	24.54**	17.59**
Qwen3-32B	0.00	-1.89	3.94	0.00	-4.17	2.75	1.18	7.81	13.54	14.12 **	9.89*	2.16	14.12**	5.56
EXAONE 4.0 32B	5.06	-9.23	5.11	18.17	19.28**	7.87	34.17**	11.60	25.70**	26.00**	17.29**	5.73*	8.51**	3.96
Qwen3-235B-A22B	13.16	8.47	3.82	26.00*	24.14**	24.58**	16.67*	-1.39	8.85	11.89**	9.09*	3.40	11.89*	3.94
Llama 3.3 70B vs R1-Distill-Llama	7.68	-8.06	-54.78**	19.57	45.45**	11.95	12.64	-1.55	17.64*	12.84*	16.29**	3.86	6.29	16.22**
DeepSeek V3 vs DeepSeek R1	16.87	5.56	8.89	18.88	33.66**	42.55**	44.73**	30.15**	22.13**	11.16**	25.16**	5.89**	7.95**	9.45**
Claude 3.7 Sonnet	-10.01	5.01	5.72	17.39*	15.50**	9.15	11.12*	-9.68	10.41*	7.85*	3.17	-1.16	0.89	-0.71

Table 10: The performance gains (%) when the reasoning is enabled for each model pair. Llama 3.3 70B stands for Llama 3.3 70B Instruct model and R1-Distill-Llama for DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI et al., 2025a). * and ** denote the statistical significance by Two-proportion Z-test, indicating p-value < 0.05 and < 0.01, respectively.

Year	**Tests**
2005	Master Craftsman Construction Equipment Maintenance, Master Craftsman Building General Work, Master Craftsman Precious Metal Processing, Master Craftsman Confectionary Making, Master Craftsman Casting, Master Craftsman Sheet-Metal & Boiler Making
2008	Master Craftsman Architectural Carpentering, Master Craftsman Surface Treatment
2010	Master Craftsman Railway Vehicles Maintenance
2014	Master Craftsman Welding
2016	Master Craftsman Steel Making
2017	Master Craftsman Metal Material, Master Craftman Metal Mould, Master Craftsman Cook
2018	Master Craftsman Gas, Master Craftsman Machinery Maintenance, Master Craftsman Plumbing, Master Craftsman Rolling, Master Craftsman Energy Management, Master Craftsman Hazardous Material, Master Craftsman Motor Vehicles Maintenance, Master Craftsman Electricity, Master Craftman Electronics, Master Craftsman Iron Making
2020	Engineer Radio Electronic Communication, Engineer Floral Design
2021	Engineer Construction Equipment, Engineer Machinery Design, Engineer Agricultural Health and Safety, Engineer Leak Nondestructive Testing, Engineer Radiation Nondestructive Testing, Engineer Biology Classification—Animal, Engineer Aquaculture, Engineer Visual Communication Design, Engineer Eddy Current Nondestructive Testing, Engineer Welding, Engineer Biomedical, Engineer Magnetic Nondestructive Testing, Engineer Electric Railway, Engineer Computer, Engineer Concrete, Engineer Explosives Handling
2022	Engineer Gas, Engineer Construction Safety, Engineer Construction Material Testing, Engineer Architecture, Engineer Building Facilities, Engineer Air-Conditioning Refrigerating Machinery, Engineer Transportation, Engineer Metal, Engineer Meteorology, Engineer Air Pollution Environmental, Engineer Urban Planning, Engineer Bioprocess, Engineer Forest, Engineer Industrial Safety, Engineer Industrial Hygiene Management, Engineer Plant Maintenance, Engineer Fire Protection System—Mechanical, Engineer Fire Protection System—Electrical, Engineer Noise & Vibration, Engineer Water Pollution Environmental, Engineer Elevator, Engineer Plant Protection, Engineer Food Processing Safety, New and Renewable Energy Equipment (Photovoltaic) Engineer, Engineer Interior Architecture, Engineer Energy Management, Engineer Greenhouse Gas Management, Engineer Organic Agriculture, Engineer Ergonomics, Engineer General Machinery, Engineer Motor Vehicles Maintenance, Engineer in Nature Environment and Ecological Restoration, Engineer Electric Work, Engineer Electricity, Engineer Computer System Application, Engineer Electronics, Engineer Information Processing, Engineer Landscape Architecture, Engineer Seeds, Engineer Cadastral Surveying, Engineer Railroad Signal Apparatus, Engineer Ultrasonic Nondestructive Testing, Engineer Livestock, Engineer Surveying Geo-Spatial Information, Engineer Penetrante Nondestructive Testing, Engineer Colorist, Engineer Civil Engineering, Engineer Soil Environment, Master Craftsman Telecommunication Apparatus, Engineer Wastes Treatment, Engineer Quality Management, Engineer Ocean Environment, Engineer Chemical Industry, Fire Investigation & Evaluation Engineer, Engineer Chemical Analysis
2023	Engineer Radio Telecommunication Equipment, Engineer Broadcasting Communication, Engineer Information Communication

Table 11: Redux Years and National Qualification Test Additions

Error Type	Examples
Ill-posed Question	Category: Political science and sociology A, B에 대한 설명으로 옳은 것만을 <보기>에서 고르면? *What is the correct explanation about A, B in <Reference>?*
Leaked Answer	Category: Ecology 산복수로에서 쌓기공작물의 높이가 3m이고, 수로깊이가 1m일 때 수로받이의 근사적 길이는? (문제 오류로 현재 복원중입니다. 보기 내용을 아시는 분들께서는 오류 신고를 통하여 보기 작성 부탁 드립니다. 정답은 3번입니다.) *What is the approximate length of the culvert if the pile is 3 meters high and the channel is 1 meter deep? (This is currently being restored due to a question error. If you know the referebce, please report the error. The correct answer is 3.)*
Notation Error	Category: Math 다항식 x2017-1을 x2-x로 나누었을 때의 나머지를 R(x)라 할 때, R(2017)의 값은? *If the remainder of the polynomial x2017-1 divided by x2-x is called R(x), what is the value of R(2017)?*
Bad Clarity	Category: Education 정신분석 상담과 행동주의 상담의 공통점에 해당하는 것은? A. 상담과정에서 과거 경험보다 미래 경험을 중시한 다 . B. 상 담 기 법 보 다 는 상 담 자 의 인 간 적 자 질 과 진 솔 한 태 도 를 중 시 한 다 . C. 인간의 행동을 인과적 관계로 해석하는 결정론적 관점을 가진 다 . D. 비합리적 신념을 인식하고 수정하는 논박 과정을 중시한 다 . *Which is a common feature of psychoanalytic counseling and behavioral counseling?* *A. Emphasizes future experiences over past experiences in the counseling process.* *B. Prioritizes the counselor's human qualities and sincerity over counseling techniques.* *C. Interprets human behavior through a deterministic perspective based on causal relationships.* *D. Focuses on the disputation process to recognize and modify irrational beliefs.*

Table 12: Examples of error types in KMMLU. Each example demonstrates a specific issue that impacts the reliability of the benchmark. Gray text represents translation of the examples in English