MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation

Weihao Xuan[1*], **Rui Yang**[2], **Heli Qi**[3], **Qingcheng Zeng**[4], **Yunze Xiao**[5], **Aosong Feng**[6], **Dairui Liu**[7], **Yun Xing**[8], **Junjue Wang**[1], **Fan Gao**[1], **Jinghui Lu**[9], **Yuang Jiang**[9], **Huitao Li**[2], **Xin Li**[2], **Kunyu Yu**[2], **Ruihai Dong**[7], **Shangding Gu**[10], **Yuekang Li**[11], **Xiaofei Xie**[12], **Felix Juefei-Xu**[13], **Foutse Khomh**[14], **Osamu Yoshie**[3], **Qingyu Chen**[6], **Douglas Teodoro**[15], **Nan Liu**[2], **Randy Goebel**[16], **Lei Ma**[1], **Edison Marrese-Taylor**[1], **Shijian Lu**[8], **Yusuke Iwasawa**[1], **Yutaka Matsuo**[1], **Irene Li**[1*]

[1]The University of Tokyo, [2]National University of Singapore, [3]Waseda University, [4]Northwestern University, [5]Carnegie Mellon University, [6]Yale University, [7]University College Dublin, [8]Nanyang Technological University, [9]Smartor LLC, [10]University of California, Berkeley, [11]University of New South Wales, [12]Singapore Management University, [13]New York University, [14]Polytechnique Montréal, [15]University of Geneva, [16]University of Alberta

weihaoxuan@g.ecc.u-tokyo.ac.jp, irene.li@weblab.t.u-tokyo.ac.jp

https://mmluprox.github.io/

Abstract

Existing large language model (LLM) evaluation benchmarks primarily focus on English, while current multilingual tasks lack parallel questions that specifically assess cross-lingual reasoning abilities. This dual limitation makes it challenging to assess LLMs' performance in the multilingual setting comprehensively. To fill this gap, we introduce ***MMLU-ProX***, a comprehensive benchmark covering 29 languages, built on an English benchmark. Each language version consists of 11,829 identical questions, enabling direct cross-lingual comparisons. Additionally, to meet efficient evaluation needs, we provide a lite version containing 658 questions per language. To ensure the high quality of ***MMLU-ProX***, we employ a rigorous development process that involves multiple powerful LLMs for translation, followed by expert review to ensure accurate expression, consistent terminology, and cultural relevance. Building on this, we systematically evaluate 36 state-of-the-art LLMs, including reasoning-enhanced and multilingual-optimized LLMs. The results reveal significant disparities in the multilingual capabilities of LLMs: While they perform well in high-resource languages, their performance declines markedly in low-resource languages, particularly for African languages. Through ***MMLU-ProX***, we aim to advance the development of more inclusive AI systems and promote equitable access to technology across global contexts.

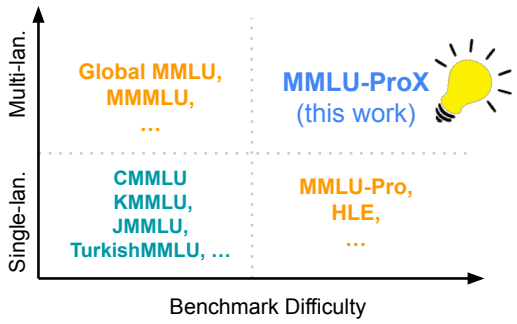

Figure 1: Selected existing benchmarks for multilingual LLM evaluation on benchmark difficulty and number of languages.

1 Introduction

The rapid development of large language models (LLMs) has significantly reshaped the field of natural language processing (NLP), with an increasing shift from predominantly English-centric systems towards multilingual understanding (Yang et al., 2025; Grattafiori et al., 2024; Aryabumi et al., 2024). As LLMs become more prevalent in global applications, the need for comprehensive multilingual evaluations becomes paramount. An effective multilingual evaluation ensures the global accessibility of LLMs, particularly benefiting users of diverse linguistic and cultural backgrounds (Poppi et al., 2024; Bang et al., 2023).

The multilingual evaluation of LLMs faces two primary challenges, as illustrated in Figure 1. First, existing benchmarks are constrained by limitations in both language coverage or translation quality. Although monolingual benchmarks such as TurkishMMLU (Yüksel et al., 2024), KMMLU (Son et al., 2024), and JMMLU (Yin et al., 2024) offer rigorous evaluation within their respective lan-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 1514–1533
November 4-9, 2025

guages, they provide limited insight for comprehensive multilingual evaluation. Broader initiatives such as Global-MMLU (Singh et al., 2025) extend coverage to 42 languages, distinguishing between culture-sensitive and culture-agnostic questions. However, the heterogeneous translation approaches pose significant challenges. The combination of professional translators, community volunteers, and Google Translate introduces quality variations that are difficult to quantify. These inconsistencies in translation quality impede objective comparison of model reasoning across languages and hinder precise diagnosis of low-resource language deficiencies. The second challenge pertains to the difficulty of the evaluation. The evolution from MMLU (Hendrycks et al., 2021a) to MMLU-Pro (Wang et al., 2024a), and Humanity's Last Exam (HLE) (Phan et al., 2025) in English benchmarking reflects the rapidly advanced reasoning capabilities of LLMs. Among those, MMLU-Pro enhances its predecessor through more complex reasoning questions, expanded answer choices, and reduction of dataset noise, offering greater discriminative power. This progression underscores the pressing need for equally challenging multilingual benchmarks that can effectively evaluate sophisticated reasoning capabilities across languages.

To address these challenges, we introduce MMLU-ProX, a novel multilingual benchmark that builds upon the challenging, reasoning-focused design of MMLU-Pro while extending its coverage to 29 typologically diverse languages. The resulting benchmark contains 11,829 questions per language in its full version, with a lite version of 658 questions available for efficient evaluation. To ensure linguistic accuracy and terminological consistency across languages, we develop a semi-automated translation agent that combines state-of-the-art (SOTA) LLMs with expert verification. This approach effectively mitigates the quality variations inherent in heterogeneous translation methods and maintains the discriminative power of MMLU-ProX in the multilingual setting.

Our primary contributions include: 1) We introduce MMLU-ProX, a multilingual benchmark for massive multitask language understanding with enhanced reasoning-focused questions across 29 languages. It enables comprehensive evaluation of LLMs' cross-lingual reasoning abilities and lays a foundation for the development of more inclusive LLMs in the future. Additionally, we engage over 30 experts to verify the data quality, with a total labor effort exceeding 400 hours. 2) We conduct systematic evaluations on MMLU-ProX and its lite version using both zero-shot and 5-shot chain-of-thought (CoT) (Wei et al., 2022) prompting across 36 latest LLMs, covering both open-source LLMs ranging from 3.8B to 671B parameters, as well as proprietary LLMs. 3) We analyze the reasoning capabilities of LLMs in the multilingual setting, revealing significant performance disparities across languages. This analysis underscores the limitations of current LLMs in global contexts, further highlighting the need to enhance global accessibility and advance fairness evaluations.

2 Related Work

Multilingual Large Language Models. The field of NLP has been profoundly transformed by multilingual LLMs, which have evolved beyond the initial English-centric paradigm to address the linguistic diversity of our world with over 7,000 languages spoken globally (Etxaniz et al., 2024). Modern LLMs are sophisticated systems built upon advanced neural architectures such as the Transformer, designed to process, comprehend, and generate text across numerous languages. Recent LLMs such as Claude 3 series (Anthropic, 2025), GPT-4 (Achiam et al., 2023), Gemini series (Google DeepMind, 2025), Qwen3 (Yang et al., 2025), and Llama 4 (Meta AI, 2025) have demonstrated remarkable multilingual capabilities. These models leverage massive pre-training datasets spanning dozens to hundreds of languages, such as the corpus used by Qwen3, encompassing 119 languages and dialects. However, research indicates persistent challenges in these systems, including the "English pivot" phenomenon (Zhong et al., 2024) where models internally process non-English inputs through English-like representations, and consistent performance gaps between high-resource and low-resource languages. Our work with MMLU-ProX specifically addresses these challenges by providing a comprehensive evaluation framework that enables direct assessment of reasoning capabilities across linguistically diverse contexts.

LLM Evaluation Benchmarks. Prior work on multilingual LLM evaluation has largely focused on breadth or translation fidelity, but often at the expense of reasoning depth or language nuance. Benchmarks like MMLU (Hendrycks et al., 2021a), TurkishMMLU (Yüksel et al., 2024) and

Dataset	Languages	Evaluation Modality	CoT	Parallel Data	Subjects	Questions
MMLU (Hendrycks et al., 2021b)	1	Multiple-choice (4)	✗	✗	57	15908
TurkishMMLU* (Yüksel et al., 2024)	1	Multiple-choice (4)	✗	✗	9	10032
KMMLU (Son et al., 2025)	1	Multiple-choice (4)	✗	✗	45	35030
XCOPA (Ponti et al., 2020a)	11	Binary choice	✗	✗	1	5500
Global-MMLU (Singh et al., 2025)	42	Multiple-choice (4)	✗	✓	57	≈ 600k
MMMLU (Hendrycks et al., 2021a)	14	Multiple-choice (4)	✗	✓	57	≈ 197k
Humanity's Last Exam (Phan et al., 2025)	1	Multiple-choice & exact match	✗	✗	2	≈ 5000
MMLU-Pro (Wang et al., 2024a)	1	Multiple-choice (10)	✓	✗	57	≈ 12k
MMLU-ProX (this work)	29	**Multiple-choice (10)**	✓	✓	57	≈ 343k

Table 1: Comparison of multilingual benchmarks with ticks (✓) and crosses (✗) indicating presence or absence of CoT and Parallel Data. *We acknowledge other MMLU datasets for various languages and randomly select TurkishMMLU as a representative example.

KMMLU (Son et al., 2025) evaluate expert reasoning tasks but are limited to a single language, while MGSM (Shi et al., 2022) and XCOPA (Ponti et al., 2020b) prioritize multilingual coverage through translated or templated questions yet restrict evaluation to narrow reasoning formats such as math problems or causal inferences. Global-MMLU (Singh et al., 2025) extends MMLU to 42 languages with human-machine hybrid translations, but it suffers from inconsistent translation quality and remains limited in reasoning difficulty. MMLU-Pro (Wang et al., 2024b) extends the original MMLU benchmark by introducing highly complex reasoning questions and more distractor options to better evaluate LLMs' reasoning depth and robustness in English. Similarly, Humanity's Last Exam (Phan et al., 2025) is a rigorous benchmark of 5,000 expert-crafted questions across diverse subjects, designed to challenge advanced AI systems and assess their progress toward expert-level reasoning, but it still remains an English-centric benchmark. While early comprehensive benchmarks such as XTREME (Ruder et al., 2021) and XGLUE (Liang et al., 2020) significantly advanced the evaluation of cross-lingual transfer, they primarily focused on traditional tasks, often assessing generalization from English training data rather than deep LLM reasoning. This landscape underscores the need for benchmarks that not only cover diverse languages but also rigorously assess complex reasoning within appropriate cultural contexts, a gap that MMLU-ProX aims to address. A detailed comparison of the aforementioned dataset is shown in Table 1. Among the selected benchmarks, MMLU-ProX fills an important gap by maintaining a balanced distribution of languages, subjects, and questions, with a focus on data parallelization and reasoning-focused features.

3 Benchmark

3.1 Overview

MMLU-ProX extends the challenging MMLU-Pro benchmark to encompass 29 typologically diverse languages: English (EN), Chinese (ZH), Japanese (JA), Korean (KO), French (FR), German (DE), Spanish (ES), Portuguese (PT), Arabic (AR), Thai (TH), Hindi (HI), Bengali (BN), Swahili (SW), Afrikaans (AF), Czech (CS), Hungarian (HU), Indonesian (ID), Italian (IT), Marathi (MR), Nepali (NE), Russian (RU), Serbian (SR), Telugu (TE), Ukrainian (UK), Urdu (UR), Vietnamese (VI), Wolof (WO), Yoruba (YO), and Zulu (ZU). MMLU-ProX benchmark maintains the high difficulty level and reasoning focus of MMLU-Pro while enabling rigorous evaluation of LLMs' cross-lingual reasoning capabilities. By carefully translating the same set of questions across all languages, MMLU-ProX facilitates direct comparison of model performance across linguistic boundaries while controlling for question difficulty.

To ensure the quality of MMLU-ProX, we implemented a multi-stage pipeline to generate the data shown in Figure 2. Initially, we hired a dedicated team to perform preliminary data curation, establishing a clean version suitable for subsequent translations. Following this, we deployed a translation agent to maintain translation quality standards. Finally, we employed a sampling methodology wherein professional translators evaluated selected samples. The results demonstrated that the generated dataset successfully passed assessment by professional human translators.

3.2 Data Curation

The data curation process comprises multiple stages. First, we identify and address duplicate or partially duplicate questions within MMLU-Pro, either eliminating or merging these instances to

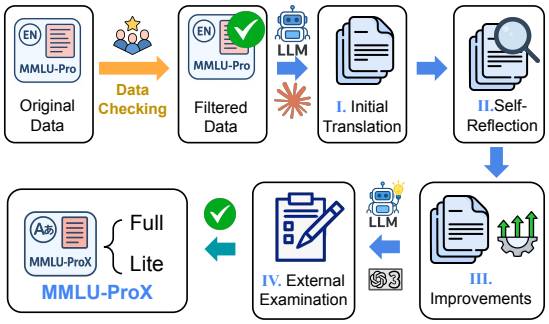

Figure 2: MMLU-ProX Data Pipeline: A rigorous four-stage process consisting of data curation, translation, external model verification, and expert review.

ensure fair evaluation without redundancy bias. Subsequently, we performed manual corrections of grammatical issues in the filtered dataset, addressing problems that include but are not limited to run-on words, incorrect hyphenation, and inconsistent symbol usage. Finally, we manually rectify inconsistencies within the questions, particularly focusing on misalignments between options and problem statements. For English data curation, we engage four interdisciplinary specialists, with a total labor investment of approximately 20 hours to correct run-on words, incorrect hyphenation, and other syntactic anomalies that could potentially confound the translation process. This curation step is critical to establish a clean baseline for our multilingual translations, as source-language errors can propagate and amplify through translation pipelines, particularly in technical and specialized domains that predominate in MMLU-Pro.

3.3 Translation Pipeline

Our data curation is followed by implementing a robust translation methodology. Based on recent machine translation evaluations (Deutsch et al., 2025; Niklaus et al., 2025), we select the SOTA Claude model, Claude Sonnet 3.7 (Anthropic, 2025) as our primary translation model. Although LLMs have shown impressive translation capabilities, we recognize the need to safeguard against potential translation errors. To address this, we develop a four-stage LLM-driven translation agent for producing MMLU-ProX:

I. Initial Translation: Claude Sonnet 3.7 performs the preliminary translations using carefully crafted prompts. These prompts emphasize maintaining accurate expression, consistent terminology across questions and options, and cultural appropriateness for target language users. The translation preserves all LaTeX notation, mathematical formulae, programming code (including variable names and comments), and currency symbols exactly as they appear in the source text. For units of measurement, we implement standard translations in target languages while maintaining precise numerical relationships and retaining all special formatting and emphasis from the original text.

II. Self-Reflection: In this stage, Claude Sonnet 3.7 performs a comprehensive review of its own translation's correspondence with the source text, generating feedback for improving the translation quality. The reflection process focuses on verifying proper noun translations and eliminating any superfluous explanations or additions. It also ensures the use of established technical terminology in the target language.

III. Improvements: Claude Sonnet 3.7 then conducts meticulous editing, incorporating feedback from the self-reflection stage. Additionally, we prompt the model to ensure the explanatory information is only included for concepts lacking direct equivalents in target languages, particularly in low-resource languages like Wolof and Yoruba. The LLM-driven process maintains strict preservation of original single quotation marks and removes any unnecessary explanations or source language terms.

IV. External Examination: To mitigate potential systematic errors from single-model biases, we employ two different LLMs for verification: OpenAI o3 for low-resource African languages (Swahili, Zulu, Yoruba, and Wolof), and GPT-4.1 for the rest. This automated verification process is designed to flag only significant discrepancies for manual review and human translation.

Appendix §F contains all translation prompts.

3.4 Expert Verification

To rigorously evaluate the benchmark quality following our translation agent implementation, we conduct comprehensive expert evaluations of the translation quality. Specifically, we randomly sample 20 items from each of the 14 disciplines and use these consistent items across all languages for evaluation. We select 15 languages and conduct expert verification, with each language evaluated by two professional translators who are native speakers of the target language and proficient in English. The total annotation effort exceeds 400 hours. Each item undergoes assessment by two high-caliber translators who rate three aspects (accuracy, fluency, and completeness) on a scale of 1 to 5. De-

tailed scoring criteria and full results can be found in Appendix §E.

Subsequently, for any category where both translators assign scores below 3 on any metric, we conduct a complete retranslation of the entire discipline of that language to ensure final averaged scores for all categories strictly exceed 4 points, an indication of accurate translation. Throughout this process, only the law category in Yoruba requires such modification; all other categories across all languages maintain average scores of 4 or above.

Table 2 presents representative translation scores, showing examples from three languages within each resource group (group criteria are in Appendix §A). The results demonstrate consistently high translation quality under expert evaluation, even for low-resource languages such as Wolof, Yoruba, and Nepali. This uniform performance across resource groups validates both the reliability of our translation pipeline and the overall quality of MMLU-ProX.

Language	Accuracy	Fluency	Completeness
High Resource			
ZH	4.70	4.84	4.92
JA	4.60	4.65	4.99
FR	4.68	4.64	4.94
Medium Resource			
KO	4.90	4.41	4.97
PT	4.79	4.77	4.99
AF	4.77	4.78	4.99
Low Resource			
WO	4.14	4.42	4.83
YO	4.06	4.56	4.95
NE	4.61	4.73	4.91

Table 2: Scores (out of 5) assigned by human translators for Accuracy, Fluency, and Completeness, grouped by language resource level (We show representatives here).

3.5 Total Cost

The development of MMLU-ProX requires substantial resource investment. Taking into account API costs for translation and testing, expert verification expenses, and computational resources, the total development cost approaches $80,000 at market rates. This investment demonstrates our commitment to creating a high-quality, reliable benchmark for advancing multilingual LLM capabilities.

4 Experiments

4.1 Setups

We evaluate a comprehensive set of 36 SOTA LLMs on MMLU-ProX across 29 linguistically diverse languages. The evaluation includes both open-weight and proprietary LLMs, representing various architectures, parameter scales, and training paradigms. The open-weight LLMs include Qwen (QwenTeam, 2025), Llama (Grattafiori et al., 2024), DeepSeek (Guo et al., 2025), Phi4 (Abdin et al., 2024), Gemma3 (GemmaTeam et al., 2025), Mistral (MistralAI, 2025), Aya (Aryabumi et al., 2024), and InternLM (Cai et al., 2024), while the proprietary LLMs comprise o4-mini, GPT-4.1 and GPT-4o. Following MMLU-Pro, we primarily employ 5-shot CoT prompting for model evaluation. All experiments were conducted on an H100 cluster. For open source models, we used vLLM for inference, while commercial models were accessed through direct API calls. Our rough estimation indicates that the unified evaluation consumed over 10,000 GPU hours.

4.2 Overall Performance

We present a comparison in Table 3, showing the CoT performance across all 29 languages and the average results of selected models, specifically the largest or best-performing model from each family (15 out of 36). We roughly group the languages by geography (stated in Appendix §B): Western Europe, South Asia, East Asia & Southeast Asia, Africa, and Eastern Europe.

Our evaluation of these LLMs reveals significant disparities in multilingual capabilities. DeepSeek-R1 demonstrates superior overall performance with an average of 75.5% across all languages, followed by GPT-4.1 (72.7%) and DeepSeek-V3 (70.5%). The performance generally correlates with model scale and architecture sophistication, with larger models typically outperforming their smaller counterparts. Among open-weight models, Qwen3-235B-Think shows exceptional capabilities, achieving SOTA results in several languages. However, there remains a substantial performance gap between high-resource and low-resource languages, with some models showing accuracy as low as 0.6% on certain African languages while achieving over 75% on Western European languages. The full evaluation for all 36 LLMs and zero-shot settings can be found in Appendix §H. We conduct a more

Language	o4-mini	GPT-4.1	GPT-4o	DeepSeek-R1	DeepSeek-V3	Mistral-3.1-24B	InternLM3-8B	Aya-32B	Phi4-14B	Llama3.3-70B	Gemma3-27B	Qwen3-32B	Qwen3-32B-Think	Qwen3-235B	Qwen3-235B-Think
Overall (AVG)	69.3	72.7	61.1	**75.5**	70.5	45.9	17.2	25.6	49.9	55.8	56.6	59.9	66.3	66.7	74.9
English (EN)	73.7	79.8	59.9	79.5	79.6	62.0	40.8	40.8	71.5	65.7	66.5	71.8	74.9	73.5	**80.7**
French (FR)	72.2	75.7	66.7	**81.3**	76.3	60.6	38.3	36.5	61.9	62.1	63.5	68.4	72.1	72.5	80.6
German (DE)	73.5	76.4	69.6	76.7	75.1	58.5	36.7	36.7	64.1	59.8	61.0	67.6	71.7	71.3	**80.4**
Spanish (ES)	74.7	77.8	68.6	80.2	76.9	59.4	36.3	35.4	59.6	61.5	63.0	68.7	72.8	73.2	**80.7**
Portuguese (PT)	74.1	77.0	67.9	78.0	75.7	60.0	36.1	30.1	61.7	61.4	63.2	69.1	72.7	73.1	**80.5**
Italian (IT)	73.9	78.2	62.9	79.9	75.9	59.6	34.7	34.5	60.2	67.0	64.4	69.4	73.5	73.7	**80.9**
Hindi (HI)	71.8	74.5	62.3	77.5	71.6	40.8	5.2	27.7	47.8	55.4	58.4	61.5	70.4	67.6	**78.7**
Bengali (BN)	70.1	72.2	62.8	66.6	69.8	32.0	3.8	14.0	34.1	50.1	55.5	57.1	66.4	67.7	**77.8**
Urdu (UR)	72.0	68.3	59.6	**76.2**	70.3	43.3	2.5	20.9	41.8	56.3	56.7	62.4	70.8	68.7	76.1
Telugu (TE)	69.1	65.9	51.3	71.9	67.6	29.3	6.4	7.2	24.1	47.9	55.9	51.0	70.3	66.7	**77.9**
Marathi (MR)	70.7	72.2	68.1	70.4	69.8	30.7	2.0	13.5	43.2	56.4	56.1	58.9	70.7	67.7	**78.5**
Nepali (NE)	71.5	74.2	61.3	**78.9**	69.3	32.9	1.5	14.5	36.0	52.8	56.8	59.7	70.7	67.8	78.1
Chinese (ZH)	72.6	75.5	64.6	**78.0**	73.9	56.5	24.2	37.4	62.3	58.4	60.4	67.0	68.7	70.5	77.4
Japanese (JA)	71.5	75.6	45.8	76.9	72.9	54.4	20.6	29.9	56.5	57.0	59.3	62.6	70.2	68.8	**77.1**
Korean (KO)	73.2	75.4	57.9	76.7	70.7	52.3	20.0	34.4	58.2	54.5	57.8	65.5	71.2	69.6	**78.3**
Vietnamese (VI)	73.4	**76.7**	70.4	76.3	75.4	53.4	5.3	30.9	57.1	65.2	61.1	68.5	72.4	71.4	72.6
Thai (TH)	72.0	75.1	66.7	**78.7**	71.2	35.4	5.5	14.9	51.7	56.0	56.7	56.1	70.4	68.8	77.1
Indonesian (ID)	73.8	75.6	66.1	**81.3**	75.8	55.5	31.6	23.1	63.9	65.5	62.6	68.5	73.4	72.5	79.9
Arabic (AR)	72.5	74.1	68.3	76.2	72.4	49.8	9.1	36.6	56.8	51.0	58.7	64.9	70.4	70.1	**78.7**
Afrikaans (AF)	73.5	77.2	65.3	**80.9**	72.9	53.3	27.6	29.7	57.8	62.7	62.0	65.9	72.4	71.1	80.6
Swahili (SW)	66.9	71.9	58.6	**75.0**	63.4	31.4	2.2	9.0	35.2	49.0	52.8	46.4	56.7	56.3	70.8
Wolof (WO)	24.1	43.2	24.3	**58.6**	47.3	17.0	0.6	1.5	8.1	28.5	8.8	26.1	26.6	26.6	36.9
Yoruba (YO)	54.9	53.4	44.3	**57.0**	47.7	13.5	0.6	3.9	23.1	31.6	32.4	25.7	18.8	40.2	49.3
Zulu (ZU)	61.2	65.0	55.3	**67.3**	53.7	17.0	2.2	14.5	11.5	33.6	40.7	17.9	35.2	46.2	46.4
Russian (RU)	62.0	71.2	62.0	76.4	74.9	59.2	26.1	36.7	65.2	61.1	62.5	68.0	69.1	72.9	**77.0**
Ukrainian (UK)	73.3	76.4	56.8	76.8	74.2	56.0	27.5	35.9	61.3	59.9	61.7	68.0	73.5	72.5	**78.8**
Serbian (SR)	72.6	76.9	70.6	**80.9**	72.9	53.9	28.9	27.4	50.7	63.0	61.7	67.2	72.3	71.1	80.2
Czech (CS)	73.5	77.5	70.1	76.8	74.7	55.1	0.0	34.5	63.2	63.8	62.6	67.7	72.8	71.8	**80.5**
Hungarian (HU)	72.6	76.6	63.0	79.1	71.4	48.7	22.2	29.1	59.4	59.7	59.8	65.9	71.1	70.1	**79.8**

Table 3: Model performance (%) on MMLU-ProX across 29 languages. Languages are grouped by geography with distinct colors. Best result per language is in **bold**. Full tables can be found in Appendix §H.

detailed analysis in the following.

4.3 Impact of Reasoning Mode in Multilingual Performance

We examine how reasoning-enhanced capabilities affect multilingual performance. Comparing reasoning-focused and standard LLMs reveals consistent performance improvements. DeepSeek-R1 outperforms DeepSeek-V3 by 5.0% on average, with larger gains in low-resource languages (Wolof: +11.3%, Yoruba: +9.3%). Similarly, Qwen3-235B with thinking mode enabled achieves superior results compared to its base performance, reaching SOTA performance on Western European languages (English: 80.7%, Spanish: 80.7%, Italian: 80.9%). These results suggest that reasoning-enhanced models better handle complex multilin-

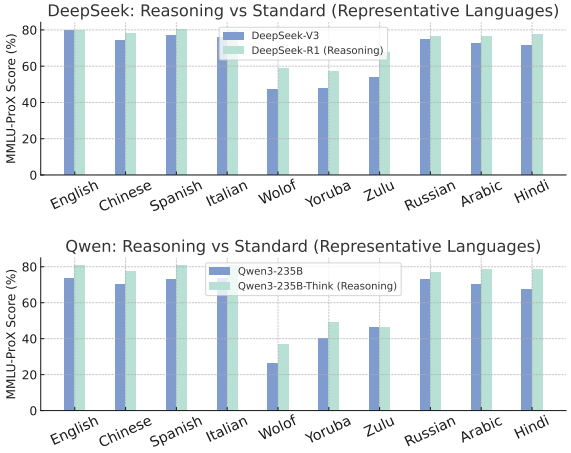

Figure 3: Comparison of reasoning-enhanced and standard models on representative languages. Top: DeepSeek-V3 vs DeepSeek-R1. Bottom: Qwen3-235B vs Qwen3-235B with thinking mode.

gual tasks, particularly in challenging linguistic contexts, indicating a promising direction for robust multilingual LLM development.

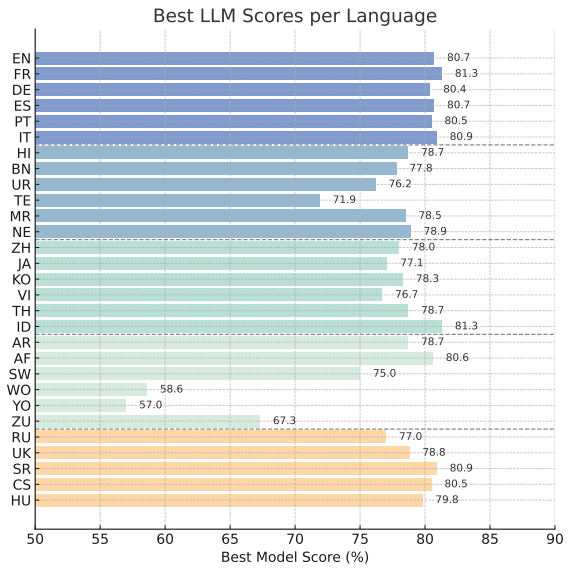

Figure 4: Best LLM scores on MMLU-ProX for each language, grouped by language family as in Table 3. The figure highlights notable performance gaps between language families, especially the advantages for well-resourced Western and Eastern European languages compared to low-resource African and some South Asian languages.

4.4 Performance across Language Groups

The results in Table 3 reveal clear performance trends across linguistic groups and model families. Western European languages consistently achieve high accuracy across all models, with top-performing models (e.g., Qwen3-235B with thinking mode exceeds 77% in every language in this group). South Asian languages show more variation: Hindi performs best within the group, while Telugu lags, highlighting challenges with Dravidian language modeling. DeepSeek-R1 and Qwen3-235B-Think stand out for their strong performance across several South Asian languages. East and Southeast Asian languages perform well overall, with Indonesian achieving 81.3% and Japanese and Korean showing stable scores, despite linguistic divergence from Indo-European languages. In contrast, African languages demonstrate the lowest performance across the board. While Arabic performs competitively, other African languages such as Wolof, Yoruba, and Zulu exhibit wide performance gaps and significantly lower scores—Wolof ranging from just 0.6% to 58.6%—highlighting the persistent limitations of current models in

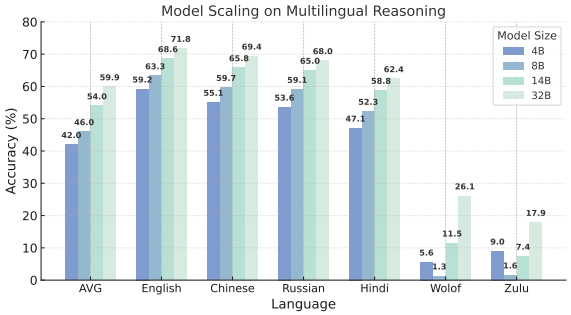

Figure 5: Performance scaling of Qwen3 dense models on selected languages.

low-resource settings. Notably, Eastern European languages also perform well, with models like DeepSeek and Qwen continuing to lead, suggesting effective adaptation to languages with linguistic similarity to Western European ones. More detailed analysis can be found in Appendix §C.

5 Analysis

In this section, we present a detailed analysis and observations on model size and prompting strategies. We then compare the full and lite versions of MMLU-ProX.

5.1 Model Size

In Figure 5, we analyze Qwen3 dense models at different scales (4B, 8B, 14B, 32B), revealing how model size affects multilingual reasoning. Performance improves consistently with scale, and the 32B model reaches 59.9% accuracy—an absolute gain of 17.9% over the 4B model on average. The largest improvement occurs from 8B to 14B (+8.0%), while gains from 4B to 8B (+4.0%) and 14B to 32B (+5.9%) are more modest. High-resource languages like English show smaller differences (12.6% from 59.2% to 71.8%), whereas low-resource languages benefit more: Wolof improves by 20.5% and Russian by 14.4%. In some African languages, such as Zulu, only the largest models show meaningful performance, suggesting a minimum model size may be required for effective multilingual capability.

These findings underscore how model scaling delivers asymmetric benefits across the linguistic spectrum, with low-resource languages typically requiring larger models to achieve even moderate performance. This differential scaling behavior highlights the importance of sufficient model capacity when developing multilingual systems intended to serve linguistically diverse user populations.

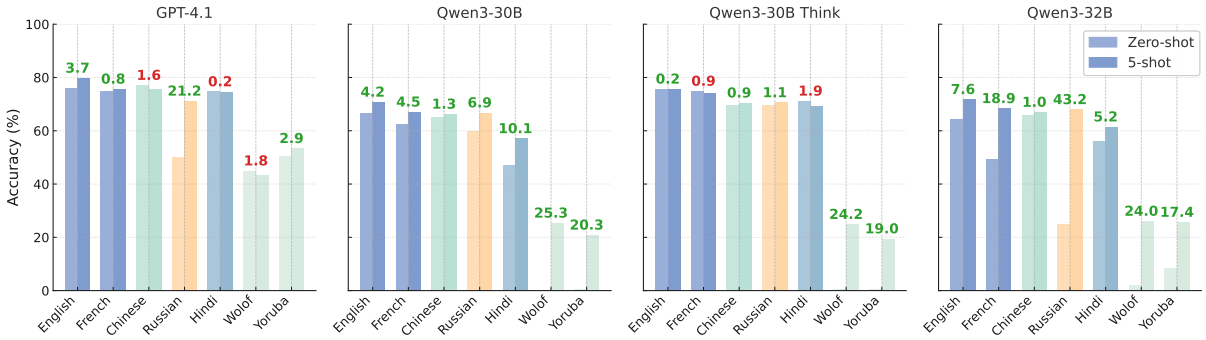

Figure 6: Performance comparison of zero-shot and 5-shot prompting across languages and models. The height of each bar represents accuracy (%). Numbers above the bar pairs indicate the absolute difference in accuracy between 5-shot and zero-shot prompting for each language-model pair. Green numbers indicate improvement, while red numbers indicate a decrease.

5.2 Prompting Strategies

For LLM evaluation, prompting strategies play a crucial role. We selected representative LLMs, including GPT-4.1 and two Qwen variants, to comprehensively evaluate the effect of different prompting strategies on multilingual reasoning capabilities. We evaluate selected languages based on resource availability and linguistic families in Figure 6.

Our analysis reveals substantial performance differences between zero-shot and 5-shot prompting across languages and model families. While 5-shot prompting generally improves performance, the magnitude of improvement varies. High-resource languages like English show modest gains (e.g., +3.7% for GPT-4.1), reflecting strong baseline reasoning abilities, whereas low-resource African languages benefit more significantly, indicating the added value of demonstrations in underrepresented languages. Reasoning-enhanced models such as Qwen3-30B in thinking mode show smaller changes between prompting styles, suggesting internalized reasoning capabilities. Additionally, language characteristics such as morphological complexity affect prompting effectiveness. These findings highlight the importance of tailoring prompting strategies to both model architecture and target language characteristics, particularly for multilingual applications targeting diverse linguistic environments.

5.3 Full and Lite Versions

To address evaluation efficiency concerns in multilingual benchmarking, we also uniformly sampled 5% of the items from each of the 14 disciplines for all 29 languages. We compare performance between the full version of MMLU-ProX (11,829 questions per language) and this lite version. Both versions include 70 validation questions used for prompt construction in few-shot evaluations, meaning actual assessments occur on 11,759 and 588 questions, respectively. As shown in Figure 7, the performance gap between both versions is remarkably small, with an average difference of only 1.14% across all evaluated models.

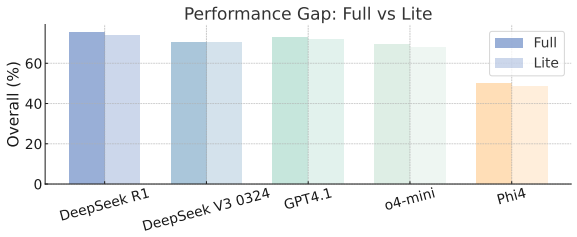

Figure 7: Comparison between MMLU-ProX full version (11,759 questions) and lite version (588 questions) across evaluated models, showing an average difference of only 1.14% while maintaining consistent relative model rankings.

Across models and language families, the lite and full versions of MMLU-ProX yield highly consistent results. Top-performing models like DeepSeek-R1 and GPT-4.1 show minimal differences between lite and full evaluations (1.5% and 1.1%, respectively), with even smaller gaps observed in models like DeepSeek-V3 (0.4%) and o4-mini (1.3%). This pattern holds across resource levels, from high-resource languages like English and French to low-resource ones like Wolof, which shows differences under 1%. Moreover, the lite version preserves the relative ranking of models almost perfectly, with DeepSeek-R1, GPT-4.1, and DeepSeek-V3 consistently outperforming others, while Phi4-14b remains the weakest. This consis-

tent ordering confirms that the lite version effectively captures the same performance patterns as the full benchmark.

6 Conclusion

We introduce MMLU-ProX, a multilingual benchmark spanning 29 diverse languages for evaluating cross-lingual capabilities of LLMs. Our semi-automatic translation approach combines LLMs with expert verification to ensure quality across languages. Additionally, we conduct a comprehensive evaluation of 36 SOTA LLMs and reveal significant performance disparities in the multilingual setting. Our work aims to promote the equitable accessibility of LLMs in the global context.

Limitations

In this work, we present MMLU-ProX, which covers 29 languages. One limitation lies in the coverage of languages due to budget constraints. While our current benchmark encompasses a diverse set of languages, expanding to include additional languages, particularly extremely low-resource ones, remains a future goal. We recognize that the existing pipeline can be extended to support such languages, and we leave this as future work.

Another limitation pertains to the expert verification of translation quality. While we engage experts to evaluate translation quality for selected languages, comprehensive expert verification across all languages and subject areas was not feasible due to resource constraints. In cases where expert evaluation was conducted, translations were assessed based on accuracy, fluency, and completeness using a 5-point Likert scale. Preliminary results indicate high overall quality, with mean scores above 4 across these dimensions. However, we acknowledge that automated translation processes may still introduce subtle errors or potential risks on the translation quality, particularly in complex or domain-specific content.

Furthermore, the current benchmark focuses solely on textual inputs and does not account for multimodal contexts, which are increasingly relevant in real-world applications. Incorporating multimodal evaluation remains an area for future exploration.

Acknowledgments

This work was supported by JST ACT-X (Grant JPMJAX24CU) and JSPS KAKENHI (Grant 24K20832). This work used supercomputers provided by the Research Institute for Information Technology, Kyushu University, through the HPCI System Research Project (Project ID: hp250092). This work is also supported by NVIDIA Academic Grant Program, Google Cloud (Gemma 3 Academic Program), and Google Academic Research Award 2025.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2025. Claude 3.7 sonnet. Large language model.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

GemmaTeam, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *ICLR*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and 1 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Meta AI. 2025. Introducing llama 4: Advancing multimodal intelligence.

MistralAI. 2025. Mistral-small-24b-instruct-2501.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill AH Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, and 1 others. 2025. Swiltra-bench: The swiss legal translation benchmark. *arXiv preprint arXiv:2503.01372*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020a. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020b. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. 2024. Towards understanding the fragility of multilingual llms against fine-tuning attacks. In *North American Chapter of the Association for Computational Linguistics*.

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

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and 1 others. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

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, and 5 others. 2025. 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. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics:*

Human Language Technologies (Volume 1: Long Papers), pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. 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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuike Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *Preprint*, arXiv:2408.10811.

A Language Categorization by Resource Availability

Following the taxonomy proposed by (Joshi et al., 2020), and referring to the resource list[1], we rank the languages from high- to low-resource as follows:
English, Chinese, Japanese, French, German, Spanish, Arabic, Korean, Portuguese, Hindi, Serbian, Hungarian, Vietnamese, Czech, Italian, Russian, Thai, Bengali, Indonesian, Ukrainian, Urdu, Afrikaans, Zulu, Swahili, Wolof, Yoruba, Telugu, Marathi, Nepali.

B Language Categorization by Geography

We primarily categorize the languages based on geography[2]. Inside each category, we rank the languages by resource availability. Below is the complete list of categories:

- **Western Europe**: English, French, German, Spanish, Portuguese, Italian
- **South Asia**: Hindi, Bengali, Urdu, Telugu, Marathi, Nepali
- **East Asia & Southeast Asia**: Chinese, Japanese, Korean, Vietnamese, Thai, Indonesian
- **Africa**: Arabic, Afrikaans, Swahili, Wolof, Yoruba, Zulu
- **Eastern Europe**: Russian, Ukrainian, Serbian, Czech, Hungarian

C Performance Patterns across Language Groups

The results in Table 3 reveal distinct patterns across linguistic families, highlighting both achievements and persistent challenges in multilingual capabilities.

As the Western European languages demonstrate consistently strong performance across all models, with scores typically ranging between 70-80% for top-performing models. In this group, Qwen3-235B with thinking achieves remarkable results, reaching 80.9% for Italian, 80.7% for both English and Spanish, and maintaining above 77% performance across all languages in this family.

South Asian languages exhibit a more nuanced performance pattern, with significant variations both across models and within the language family. Hindi consistently leads this group with scores ranging from 58.4% to 78.7%, while related languages like Bengali and Marathi show slightly lower but stable performance patterns. Telugu, representing the Dravidian family, generally shows lower performance across models, highlighting potential challenges in handling its distinct linguistic features. DeepSeek-R1 and Qwen3-235B-Think demonstrate particularly strong capabilities in this group, consistently achieving scores above 75% for several languages.

East Asian & Southeast Asian languages present an interesting case of high performance with model-specific variations. Chinese shows notable fluctuations across models (53.4-75.5%), while Japanese and Korean demonstrate more consistent performance patterns. Southeast Asian languages perform remarkably well, with Indonesian achieving 81.3% with DeepSeek-R1. This success suggests effective handling of these diverse linguistic structures by modern LLMs, despite the significant typological differences from Western languages.

African languages reveal the most pronounced performance disparities, underscoring critical challenges in multilingual AI development. While Arabic achieves competitive scores (up to 78.7% with Qwen3-235B-Think), other African languages show substantially lower performance. Wolof presents the most challenging case, with scores ranging dramatically from 0.6% to 58.6%, highlighting severe resource limitations. Similar patterns emerge for Yoruba (3.9-57.0%) and Zulu (11.5-67.3%), though with slightly better performance than Wolof. These stark contrasts emphasize the ongoing need for improved model capabilities in low-resource languages.

Notably, Eastern European languages exhibit comparable performance. Similarly, models from the DeepSeek and Qwen families continue to perform strongly, with Qwen3-235B-Think achieving over 77%. These strong performances suggest effective adaptation to these languages, likely due to their shared linguistic structures with Western European languages.

[1] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt
[2] https://www.cia.gov/the-world-factbook/field/languages/

D Translation Pipeline Analysis

For our translation agent evaluation, we compared two competitors in English-to-Japanese translation: reasoning-based translation and human translators. Using the same samples as in Section 3.4, we conducted translations using these two methods and employed professional translators to score them using our 5-point scale. The results are presented in Table 4. Our findings reveal that the reasoning-based method achieves translation quality only marginally inferior to our translation agent. However, compared to our agent-based method, reasoning-based translation consumes significantly more tokens, causing higher translation costs. As for native-speaking translators, their translation quality, particularly accuracy, proved inferior to LLM-based translation when handling content requiring multidisciplinary expertise. These results demonstrate the effectiveness of our comprehensive framework and further validate the quality of MMLU-ProX data.

Method	Accuracy	Fluency	Completeness
Agent-based Translation (Ours)	**4.60**	**4.65**	**4.99**
Reasoning-based Translation	4.56	4.21	**4.99**
Native-Speaking Translator	4.24	4.14	4.56

Table 4: Scores (out of 5) assigned by human translators for Accuracy, Fluency, and Completeness, grouped by language resource level.

E Expert Verification Guidance

We ensured that all expert annotators were compensated at rates above the minimum hourly wage in their respective countries. Evaluation Criteria for Expert Rating of Machine Translation Results:

1. Accuracy (1-5):

- **5 (Highly Accurate):**
  - All key terms and concepts are translated correctly with no errors.
  - Every technical term corresponds precisely to the original text, with no mistranslations or incorrect word choices.
  - The most appropriate and professional terminology in the target language is used.
  - Expressions align with commonly used terminology in professional or technical contexts.
- **4 (Accurate):**
  - Most terms and concepts are translated correctly, with only a few minor errors that do not affect overall comprehension.
  - Some terms may be slightly imprecise, but the translation remains generally accurate.
  - Uses appropriate terminology in the target language in most cases.
  - A few terms may be simplified but remain understandable within the intended domain.
- **3 (Moderately Accurate):**
  - Key terms and concepts are mostly correct but contain some errors that may cause partial misunderstandings.
  - Some critical terms are inaccurately translated, requiring the reader to infer the intended meaning.
  - Slight deviations in the use of target-language terminology.
  - Occasionally uses uncommon or outdated terms.
- **2 (Somewhat Inaccurate):**
  - Many key terms and concepts are mistranslated, significantly affecting comprehension.

  - Important concepts are incorrectly translated, leading to potential misunderstandings of the original text.
  - Uses incorrect or inappropriate terminology in the target language.
  - Terminology is inconsistent, reducing the text's professionalism.

- **1 (Inaccurate):**
  - Frequent and severe mistranslations of key terms and concepts, failing to convey the original meaning.
  - Most of the content does not match the original text.
  - Lacks proper use of target-language terminology.
  - Terminology is chaotic, possibly using irrelevant or incorrect vocabulary entirely.

2. Fluency (1–5):

- **5 (Highly Fluent):**
  - The target-language expression is natural and smooth, making it effortless to read.
  - The language style is refined and appropriate for professional or formal contexts.
  - The sentence structure fully adheres to natural conventions in the target language, with no grammatical or lexical errors.

- **4 (Fluent):**
  - The target-language expression is generally natural, with only minor linguistic imperfections that do not affect comprehension.
  - Some sentences may sound slightly stiff.
  - Sentence structures mostly conform to target-language norms, with very few grammatical errors.

- **3 (Moderately Fluent):**
  - The target-language expression is somewhat unnatural, requiring the reader to adjust their understanding slightly.
  - Some inappropriate word choices or rigid sentence structures are present.
  - Sentence structures are mostly correct, but some grammatical errors exist.

- **2 (Somewhat Unnatural):**
  - The target-language expression lacks fluency, making it difficult to read smoothly.
  - Sentence transitions are awkward, and logical connections are unclear.
  - Many structural issues exist, with frequent grammatical errors.

- **1 (Not Fluent):**
  - The target-language expression is highly unnatural or difficult to understand.
  - Literal translation is evident, lacking natural phrasing in the target language.
  - The sentence structure is disorganized, with severe grammatical mistakes, making the text unreadable.

3. Completeness (1–5):

- **5 (Fully Complete):**
  - The full meaning of the original text is retained with no omissions or additions.
  - All details, data, and annotations are accurately conveyed.
  - The translation maintains the same length and depth as the original text.

- **4 (Complete):**
  - The primary meaning of the original text is retained, with only a few minor details omitted or slightly unclear.
  - Some less critical information may be left out.
  - The translation generally corresponds to the original content.
- **3 (Moderately Complete):**
  - Most of the original meaning is conveyed, but some information is missing or added.
  - Important details may be overlooked.
  - The translation differs from the original in certain aspects, requiring readers to infer some content.
- **2 (Somewhat Incomplete):**
  - The core information from the original text is not fully conveyed, with noticeable omissions or unnecessary additions.
  - Potential inclusion of unrelated information.
  - The translation does not fully correspond to the original, affecting comprehension.
- **1 (Incomplete):**
  - Significant omissions or added incorrect information prevent an accurate reflection of the original text.
  - Important sections or sentences are missing.
  - The translation deviates heavily from the original, making it difficult to understand the intended meaning.

Scoring Examples:

- **Accuracy Example:**
  If "bachelor's degree" is mistranslated as "single man's degree," points should be deducted in the accuracy category.
- **Fluency Example:**
  If the sentence structure follows target-language norms but the word choice is slightly unnatural, a score of 4 may be appropriate.
- **Completeness Example:**
  If the translated text omits the methodology section from the original, it should receive a lower score in completeness.

We perform expert verification in 15 selected languages, the full results are shown in Table 5.

Language	Accuracy	Fluency	Completeness
High Resource			
ZH	4.70	4.84	4.92
JA	4.60	4.65	4.99
FR	4.68	4.64	4.94
DE	4.52	4.48	4.64
ES	4.59	4.58	4.84
Medium Resource			
KO	4.90	4.41	4.97
PT	4.79	4.77	4.99
AF	4.77	4.78	4.99
Low Resource			
ZU	4.20	4.62	4.97
SW	4.36	4.70	4.86
WO	4.14	4.42	4.83
YO	4.06	4.56	4.95
TE	4.60	4.74	4.97
MR	4.51	4.72	4.74
NE	4.61	4.73	4.91

Table 5: Expert verification scores on 15 languages for Accuracy, Fluency, and Completeness, grouped by language resource level.

F Translation Prompts

initial_translation

System Message: You are a professional translator specializing in accurate translation of technical and academic content from {source_lang} to {target_lang}.

Your task is to translate assessment questions in the {category} field while:
1. Preserving technical accuracy and terminology
2. Ensuring cultural appropriateness for {target_lang} speakers
3. Keeping terminology consistent throughout questions and options
4. Preserving all LaTeX notation, mathematical formulas, and programming code exactly as they appear (do not translate content inside LaTeX delimiters or code blocks, including variable names, function names, and comments)
5. Preserving all currency symbols ($) exactly as they appear in the original text, without converting to local currency units
6. For units of measurement: Use the conventional translations in the target language while preserving the exact numerical values and relationships
7. Preserving any special formatting or emphasis in the original text

Please translate the following {category} assessment question and options:
<SOURCE_TEXT>
{source_text}
</SOURCE_TEXT>

Output:
Only provide the {target_lang} translation for the above text. Do not include any explanations or text apart from the translation.
Different options are separated by newline characters(\n).
The number of options in the output must match the input exactly. Do not skip or combine any options.
Return the translation in the following JSON format, with keys "question" and "options", where the value of "options" is a dictionary with keys option1, option2, option3, etc. All JSON keys must remain in English exactly as shown and only translate the content inside square brackets:

<TRANSLATION>
{{
"question": "[translation of question]",
"options": {{
"option1": "[translation of option1]",
"option2": "[translation of option2]",

"option3": "[translation of option3]",
...
}}
}}
</TRANSLATION>

F.1 Self-Reflection Prompt

self_reflection

System Message: You are a {category} translation expert, specializing in translation from {source_lang} to {target_lang}.

Task Description:
Carefully review the source text and its translation from {source_lang} to {target_lang}, and then provide constructive suggestions in English.

Requirements:
1. Do not add, remove, or explain any information.
2. Make sure retain the original format for specialized information, e.g., anonymous information.
3. Identify any instances where proper nouns remain untranslated or where the translation contains unnecessary explanations, parenthetical original terms, or additions from {source_lang}.
4. Examine whether any technical terms, subject-specific concepts, or other specialized vocabulary have been left in {source_lang} instead of using their established {target_lang} equivalents.
5. Verify that currency symbols, mathematical operators, and measurement units remain exactly as they appear in {source_lang} text. These symbols should not be converted to their written form in {target_lang}.
6. Check that no additional symbols or written representations have been added to options where they did not exist in {source_lang} text.

Input:
<SOURCE_TEXT>
{source_text}
</SOURCE_TEXT>

<INITIAL_TRANSLATION>
{initial_trans}
</INITIAL_TRANSLATION>

Output:
<SUGGESTIONS>
[Your suggestions here]
</SUGGESTIONS>

F.2 Translation Improvement Prompt

improve_translation

System Message: You are a {category} translation expert, specializing in translation from {source_lang} to {target_lang}.

Task Description:
Carefully review and edit the {category} translation from {source_lang} to {target_lang}, incorporating the expert feedback.

Requirements:
1. Do not explain any information.
2. Strictly keep the single quotes in the original text and do not add new single and double quotes.
3. Remove unnecessary explanations or original terms from {source_lang} if present in the translation.

Input:
<SOURCE_TEXT>
{source_text}
</SOURCE_TEXT>

<INITIAL_TRANSLATION>
{initial_trans}
</INITIAL_TRANSLATION>

<EXPERT_SUGGESTIONS>
{reflection}
</EXPERT_SUGGESTIONS>

Output:
Only provide the improved translation. Do not include any explanations or text apart from the translation.
Different options are separated by newline characters(\n).
The number of options in the output must match the input exactly. Do not skip or combine any options.
Return the translation in the following JSON format, with keys "question" and "options", where the value of "options" is a dictionary with keys option1, option2, option3, etc. All JSON keys must remain in English exactly as shown and only translate the content inside square brackets:

<IMPROVED_TRANSLATION>
{{
"question": "[improved translation of question]",
"options": {{
"option1": "[improved translation of option1]",
"option2": "[improved translation of option2]",
"option3": "[improved translation of option3]",
...
}}
}}
</IMPROVED_TRANSLATION>

G Generative AI Statement

Large language models were utilized to facilitate aspects of the dataset creation in this project. Specifically, Claude Sonnet 3.7 was employed to assist with the translation of benchmark content into multiple languages, and GPT-4.1 and o3 were used to provide external verification of translation quality.

H Detailed Evaluation Results

We put more results in Table 6, Table 7, Table 8, and Table 9.

Model	Avg	EN	FR	DE	ES	PT	IT	HI	BN	UR	TE	MR	NE	ZH	JA	KO	VI	TH	ID	AR	AF	SW	WO	YO	ZU	RU	UK	SR	CS	HU
o4-mini	69.3	73.7	72.2	73.5	74.7	74.1	73.9	71.8	70.1	72.0	69.1	70.7	71.5	72.6	71.5	73.2	73.4	72.0	73.8	72.5	73.5	66.9	24.1	54.9	61.2	62.0	73.3	72.6	73.5	72.6
GPT-4o	61.1	59.9	66.7	69.6	68.6	67.9	62.9	62.3	62.8	59.6	51.3	68.1	61.3	64.6	45.8	57.9	70.4	66.7	66.1	68.3	65.3	58.6	24.3	44.3	55.3	62.0	56.8	70.6	70.1	63.0
GPT-4.1	72.7	79.8	75.7	76.4	77.8	77.0	78.2	74.5	72.2	68.3	65.9	72.2	74.2	75.5	75.6	75.4	**76.7**	75.1	75.6	74.1	77.2	71.9	43.2	53.4	65.0	71.2	76.4	76.9	77.5	76.6
DeepSeek-R1-7B	17.2	28.9	31.8	31.9	38.7	37.3	39.6	4.7	0.3	3.3	2.5	1.9	5.6	33.5	2.6	3.3	33.1	10.8	40.4	28.8	23.1	0.9	0.3	0.0	0.3	18.4	24.2	26.3	25.3	0.7
DeepSeek-R1-8B	18.4	36.2	28.5	33.6	33.9	25.4	31.4	21.4	0.0	4.7	0.0	21.5	8.2	24.5	21.3	10.6	26.9	21.1	34.2	23.2	20.1	3.2	0.5	0.4	0.6	26.9	26.1	21.9	21.9	6.4
DeepSeek-R1-14B	33.7	59.2	53.6	48.1	58.8	50.9	55.0	21.0	28.7	4.3	4.4	4.1	7.1	33.1	45.8	21.0	55.8	57.5	53.3	53.2	22.0	16.9	22.0	0.2	0.1	20.9	59.7	33.1	42.5	11.2
DeepSeek-R1-32B	45.2	71.1	61.9	58.3	67.0	58.3	61.5	43.7	42.5	39.2	7.7	21.4	34.9	55.2	57.9	43.7	62.9	58.8	62.7	54.4	39.2	9.3	40.9	0.1	0.5	60.3	61.6	64.9	57.8	38.6
DeepSeek-R1-70B	58.7	70.3	73.7	72.5	73.7	70.6	73.6	62.0	49.3	56.9	41.2	48.4	44.0	53.4	59.8	47.4	69.3	60.0	71.5	62.6	73.1	62.0	12.2	26.3	46.7	52.7	57.5	67.1	71.7	71.5
DeepSeek-R1	**75.5**	79.5	**81.3**	76.7	80.2	78.0	79.9	77.5	66.6	**76.2**	71.9	70.4	**78.9**	**78.0**	76.9	76.7	76.3	**78.7**	**81.3**	76.2	**80.9**	**75.0**	**58.6**	**57.0**	**67.3**	76.4	76.8	**80.9**	76.8	79.1
DeepSeek-V3	70.5	79.6	76.3	75.1	76.9	75.7	75.9	71.6	69.8	70.3	67.6	69.8	69.3	73.9	72.9	70.7	75.4	71.2	75.8	72.4	72.9	63.4	47.3	47.7	53.7	74.9	74.2	72.9	74.7	71.4
Mistral-Small-24B	42.8	61.2	58.2	56.5	57.7	57.8	57.7	32.0	30.8	37.6	25.7	27.2	27.8	60.7	51.8	50.3	51.1	36.1	52.6	30.2	51.8	26.4	17.1	10.2	14.8	56.8	54.9	53.1	52.4	47.9
Mistral-Small-3.1-24B	45.9	62.0	60.6	58.5	59.4	60.0	59.6	40.8	32.0	43.3	29.3	30.7	32.9	56.5	54.4	52.3	53.4	35.4	55.5	49.8	53.3	31.4	17.0	13.5	17.0	59.2	56.0	53.9	55.1	48.7
InternLM3-8B	17.2	40.8	38.3	36.7	36.3	36.1	34.7	5.2	3.8	2.5	6.4	2.0	1.5	24.2	20.6	20.0	5.3	5.5	31.6	9.1	27.6	2.2	0.6	0.6	2.2	26.1	27.5	28.9	0.0	22.2
Aya-23-8B	13.7	34.4	18.4	20.6	19.1	11.4	18.0	11.8	3.4	4.7	3.4	4.2	4.7	27.5	18.0	20.5	16.1	5.5	11.1	19.6	19.4	5.3	3.0	4.1	3.2	19.2	17.6	14.4	19.0	18.8
Aya-23-32B	25.6	40.8	36.5	36.7	35.4	30.1	34.5	27.7	14.0	20.9	7.2	13.5	14.5	37.4	29.9	34.4	30.9	14.9	23.1	36.6	29.7	9.0	1.5	3.9	14.5	36.7	35.9	27.4	34.5	29.1
Phi4-mini-3.8B	21.6	51.6	37.3	36.8	37.9	37.7	35.3	13.0	11.7	14.0	14.9	12.7	8.0	35.3	23.9	14.3	24.6	17.6	14.0	27.4	18.0	13.7	3.3	3.4	2.1	32.6	22.8	24.4	21.1	16.8
Phi4-14B	49.9	71.5	61.9	64.1	59.6	61.7	60.2	47.8	34.1	41.8	24.1	43.2	36.0	62.3	56.5	58.2	57.1	51.7	63.9	56.8	57.8	35.2	8.1	23.1	11.5	65.2	61.3	50.7	63.2	59.4
Llama3.1-8B	22.9	45.2	31.2	32.2	20.7	37.8	34.8	20.5	18.8	14.2	14.2	19.7	17.3	31.9	24.9	24.0	31.3	27.9	22.6	13.4	27.9	15.1	0.4	6.4	4.5	28.8	24.1	27.1	23.1	25.5
Llama3.1-70B	49.3	62.1	58.2	55.4	58.3	57.3	59.2	48.9	45.7	49.1	41.8	46.0	44.6	54.6	51.7	49.5	55.5	52.8	51.3	48.6	57.1	44.6	26.6	24.3	23.8	49.2	50.3	56.1	55.9	50.5
Llama3.3-70B	55.8	65.7	62.1	59.8	61.5	61.4	67.0	55.4	50.1	56.3	47.9	56.4	52.8	58.4	57.0	54.5	65.2	56.0	65.5	51.0	62.7	49.0	28.5	31.6	33.6	61.1	59.9	63.0	63.8	59.7
Llama4-Scout	51.3	71.2	63.0	66.6	62.8	67.4	57.8	40.1	47.6	40.3	42.4	49.3	50.0	40.2	49.6	42.0	51.1	52.4	58.4	49.4	58.2	49.0	14.2	32.2	41.2	61.4	58.3	53.0	57.5	60.2
Gemma3-4B	28.1	40.8	35.3	34.3	36.4	37.1	36.4	28.6	27.1	18.1	25.4	24.9	22.2	33.4	31.4	28.3	32.4	26.9	33.5	29.7	28.6	18.0	7.0	12.7	9.8	34.7	32.9	32.1	33.5	24.9
Gemma3-12B	48.5	59.8	55.6	54.3	55.8	55.5	56.3	49.6	46.0	43.6	46.7	47.7	47.2	51.7	50.8	49.1	54.5	49.7	55.4	49.9	50.7	44.0	9.4	25.0	32.0	54.0	53.9	53.1	54.8	50.5
Gemma3-27B	56.6	66.5	63.5	61.0	63.0	63.2	64.4	58.4	55.5	56.7	55.9	56.1	56.8	60.4	59.3	57.8	61.1	56.7	62.6	58.7	62.0	52.8	8.8	32.4	40.7	62.5	61.7	61.7	62.6	59.8
Qwen2.5-3B	24.8	43.9	35.6	34.3	34.7	31.8	32.7	21.3	14.2	24.5	19.4	23.3	9.2	34.1	30.1	28.4	29.9	27.4	31.6	28.6	28.9	6.6	1.1	10.6	5.4	26.2	28.5	23.8	28.8	24.3
Qwen2.5-7B	37.6	57.5	48.9	46.9	49.3	46.1	49.1	34.0	32.2	25.7	23.4	29.5	27.3	50.5	43.6	41.5	46.4	39.6	46.6	40.2	41.1	23.2	11.0	21.1	13.3	46.3	42.9	39.7	42.3	31.3
Qwen2.5-14B	47.7	65.5	59.1	58.5	59.6	59.0	59.8	43.4	41.5	41.1	33.6	39.2	38.3	59.9	54.7	53.5	57.1	50.3	55.9	51.0	53.3	24.2	26.3	25.9	8.6	59.1	54.0	50.1	53.7	46.0
Qwen2.5-32B	54.2	68.7	65.2	63.4	65.1	64.7	65.9	49.1	48.1	50.7	38.9	49.0	48.7	62.6	61.7	59.1	62.4	56.8	64.2	58.5	59.8	35.0	27.2	27.3	19.4	64.0	61.6	59.8	60.4	56.1
Qwen2.5-72B	58.8	70.3	67.1	65.9	66.5	66.6	68.8	58.0	57.6	59.1	47.9	56.4	57.3	65.9	63.4	62.1	66.4	60.1	68.2	62.1	65.4	40.1	29.0	27.9	25.9	68.1	66.1	65.4	66.8	62.3
Qwen3-4B	42.0	59.2	54.5	52.6	54.6	54.3	55.1	39.3	35.5	34.3	24.4	32.6	35.7	54.3	49.7	46.7	52.5	47.1	52.9	44.5	48.7	16.2	5.6	8.9	9.0	53.6	51.2	49.1	49.8	45.8
Qwen3-4B-Think	53.8	65.2	62.9	62.9	63.3	63.4	63.1	56.9	53.8	48.9	52.3	55.3	55.1	54.3	61.3	58.7	62.3	58.9	62.6	58.0	59.7	30.4	19.3	16.7	17.3	53.8	61.2	60.2	61.6	60.0
Qwen3-8B	46.0	63.3	59.8	57.8	59.8	58.7	59.7	47.0	44.3	44.6	21.1	37.2	43.3	58.5	54.8	52.9	57.3	52.3	58.9	52.1	54.6	6.9	1.3	8.4	1.6	59.1	55.6	54.7	56.4	51.6
Qwen3-8B-Think	58.8	70.6	68.3	67.9	68.4	68.0	68.6	63.2	61.3	60.7	56.8	63.0	60.4	60.0	66.7	63.7	58.9	64.9	68.1	64.4	66.7	31.5	19.9	24.1	14.2	59.8	66.9	66.7	67.1	65.4
Qwen3-14B	54.0	68.6	65.2	64.4	65.7	65.6	65.8	53.7	51.7	52.1	46.5	50.9	50.9	64.2	61.8	58.9	63.7	58.8	64.5	58.2	60.2	29.0	11.5	16.8	7.4	65.0	62.3	62.2	62.2	58.5
Qwen3-14B-Think	65.4	74.7	72.2	71.4	72.2	71.6	73.0	67.9	66.8	68.0	64.6	66.9	67.3	66.5	70.8	70.0	71.4	69.8	72.3	69.1	71.3	48.0	28.3	32.5	32.3	72.2	71.3	71.6	71.1	69.8
Qwen3-30B-A3B	57.0	70.8	67.1	65.0	66.9	66.3	67.5	57.1	53.8	54.3	50.5	53.8	54.0	66.3	63.9	60.9	65.1	60.7	65.9	60.5	62.4	29.5	25.4	20.9	23.1	66.8	64.3	64.1	64.9	61.9
Qwen3-30B-A3B-Think	64.8	75.7	74.0	73.4	74.0	73.8	72.0	69.3	68.6	69.2	69.4	69.8	69.0	70.5	72.1	71.4	38.9	70.3	74.0	70.6	72.4	49.6	24.9	19.5	26.2	70.6	72.8	72.7	73.6	72.2
QwQ-32B	63.2	75.9	75.7	74.1	76.0	75.7	75.6	71.3	70.0	59.8	58.2	59.2	68.4	66.6	66.3	65.6	72.7	72.5	70.6	62.3	73.5	45.8	8.4	19.4	21.7	74.4	70.5	64.4	73.5	65.8
Qwen3-32B	59.9	71.8	68.4	67.6	68.7	69.1	69.4	61.5	57.1	62.4	51.0	58.9	59.7	67.0	62.6	65.5	68.5	56.1	68.5	64.9	65.9	46.4	26.1	25.7	17.9	68.0	68.0	67.2	67.7	65.9
Qwen3-32B-Think	66.3	74.9	72.1	71.7	72.8	72.7	73.5	70.4	66.4	70.8	70.3	70.7	70.7	68.7	70.2	71.2	72.4	70.4	73.4	70.4	72.4	56.7	26.6	18.8	35.2	69.1	73.5	72.3	72.8	71.1
Qwen3-235B	66.7	73.5	72.5	71.3	73.2	73.1	73.7	67.6	67.7	68.7	66.7	67.7	67.8	70.5	68.8	69.6	71.4	68.8	72.5	70.1	71.1	56.3	26.6	40.2	46.2	72.9	72.5	71.1	71.8	70.1
Qwen3-235B-Think	74.9	**80.7**	80.6	**80.4**	**80.7**	**80.5**	**80.9**	**78.7**	**77.8**	76.1	**77.9**	**78.5**	78.1	77.4	**77.1**	**78.3**	72.6	77.1	79.9	**78.7**	80.6	70.8	36.9	49.3	46.4	**77.0**	**78.8**	80.2	**80.5**	**79.8**

Table 6: MMLU-ProX 5-shot Results.

Model	Avg	EN	FR	DE	ES	PT	IT	HI	BN	UR	TE	MR	NE	ZH	JA	KO	VI	TH	ID	AR	AF	SW	WO	YO	ZU	RU	UK	SR	CS	HU
GPT-4.1	**71.1**	**76.1**	**74.9**	**76.1**	73.8	**77.3**	75.5	**74.7**	**72.9**	**72.6**	60.3	**74.0**	**74.2**	**77.1**	**73.9**	**75.1**	**76.4**	**74.1**	67.4	**75.1**	**76.2**	**71.1**	**45.0**	**50.5**	**65.3**	50.0	**76.1**	**76.2**	**75.9**	**75.4**
Mistral-31-24B	50.4	67.5	63.9	62.5	64.1	61.7	64.9	51.1	42.6	51.0	42.1	41.9	39.6	61.1	56.2	55.6	53.8	42.9	60.9	53.6	54.2	39.2	4.8	7.4	20.6	63.1	60.8	60.8	57.2	56.4
Qwen2.5-72B	58.8	72.3	68.8	67.7	68.2	68.7	69.6	59.9	57.1	59.7	37.9	55.1	57.2	66.0	64.5	62.1	67.5	63.4	68.4	64.0	65.9	41.9	19.8	24.3	24.0	68.9	66.5	66.9	67.1	60.9
Qwen3-30B-A3B	49.1	66.6	62.6	66.1	67.4	67.0	68.3	47.0	40.9	40.6	31.8	39.0	35.0	65.0	63.5	62.4	63.6	45.5	48.4	59.8	55.4	21.3	0.1	0.6	8.3	59.9	52.8	62.3	64.6	58.1
Qwen3-30B-A3B-Think	65.0	75.5	**74.9**	74.1	**75.0**	74.6	**75.8**	71.2	70.4	71.4	**69.4**	71.2	66.2	69.6	73.6	73.0	73.8	71.6	**75.0**	71.7	73.9	55.8	0.7	0.5	13.3	**69.5**	73.8	74.0	74.8	71.0
Qwen3-32B	51.3	64.2	49.5	53.8	69.5	67.9	68.7	56.3	44.3	52.2	29.0	52.9	53.4	66.0	58.8	66.1	67.3	44.8	57.7	63.9	57.2	27.4	2.1	8.3	23.3	24.8	58.1	68.2	69.1	63.9
Qwen3-32B-Think	61.9	71.1	70.4	69.1	71.8	71.5	71.5	67.0	65.4	67.6	56.3	68.4	63.3	67.3	68.5	69.9	69.7	66.6	70.9	68.0	69.5	57.2	0.8	10.0	19.1	63.6	70.8	71.2	72.1	65.2

Table 7: MMLU-ProX zero-shot Results.

Model	Avg	EN	FR	DE	ES	PT	IT	HI	BN	UR	TE	MR	NE	ZH	JA	KO	VI	TH	ID	AR	AF	SW	WO	YO	ZU	RU	UK	SR	CS	HU
GPT-4.1	71.6	78.7	75.2	73.6	77.6	74.3	79.8	73.0	70.4	66.8	65.5	69.4	72.6	73.0	73.8	74.7	**76.4**	74.3	75.9	73.0	76.9	71.1	41.5	51.7	62.9	69.9	75.9	76.0	76.2	76.5
o4-mini	68.0	73.3	72.4	72.4	73.1	73.1	73.6	69.6	69.7	69.7	67.7	69.0	68.4	70.1	68.2	71.6	73.5	69.6	71.4	73.0	72.3	66.8	25.0	51.9	61.7	58.8	71.9	73.5	71.9	68.4
DeepSeek-R1	74.0	78.9	80.6	72.1	**80.4**	77.2	78.7	75.0	65.3	74.0	70.7	68.7	76.2	75.5	75.5	73.1	75.0	**79.3**	78.6	74.8	77.2	**71.4**	**59.5**	**54.3**	**67.7**	77.2	76.4	**79.4**	75.2	78.1
DeepSeek-V3	70.1	79.6	74.7	74.8	76.7	76.4	76.9	70.2	68.0	69.2	66.2	68.4	67.5	74.5	73.0	68.5	75.2	69.7	77.7	71.9	71.8	64.8	45.2	46.8	54.6	75.9	75.2	72.1	75.0	72.8
Phi4-14B	48.5	69.0	61.1	61.1	56.8	62.6	60.5	47.6	34.4	41.8	23.3	42.9	33.8	57.8	53.9	55.6	56.1	50.9	61.1	57.5	54.4	35.9	6.5	21.8	10.9	61.7	57.1	50.9	62.8	57.7
Gemma3-27B	56.0	65.6	64.8	60.5	60.4	59.9	62.8	59.4	56.1	58.2	54.6	56.8	55.8	58.7	56.6	55.6	60.9	55.8	63.6	58.0	61.6	51.9	8.7	32.5	38.4	63.6	60.5	61.9	63.8	57.1
Llama3.3-70B	55.7	68.0	65.6	63.6	64.5	65.1	65.0	53.9	50.7	52.7	46.6	54.9	49.5	60.0	59.0	56.3	61.2	56.0	62.4	54.8	63.4	55.1	28.1	28.4	32.1	59.2	57.7	60.0	63.3	58.3
Llama4-Scout	49.8	69.0	65.1	62.9	58.5	64.8	56.3	36.7	46.6	38.6	40.1	48.8	49.1	38.6	47.8	41.3	48.0	49.3	56.6	48.8	57.0	49.0	15.0	30.8	40.1	60.7	56.3	52.9	56.3	59.0
Qwen2.5-72B	59.8	70.4	69.0	68.4	69.4	68.4	68.5	59.4	59.4	61.7	49.8	56.0	57.7	67.3	65.0	65.1	66.7	62.2	66.7	62.9	64.5	41.8	28.2	28.6	25.5	70.7	67.9	63.9	66.8	61.6
Qwen3-30B-A3B	56.0	70.4	65.8	61.6	65.3	65.6	68.0	57.0	54.6	53.1	53.4	52.4	49.1	65.3	61.9	60.5	64.1	60.0	62.4	58.3	61.1	28.7	25.7	20.2	24.7	65.3	62.1	62.4	64.1	60.9
Qwen3-30B-A3B-Think	63.8	75.2	75.3	71.8	71.4	73.5	70.4	67.0	66.3	69.9	67.2	68.2	67.2	67.3	69.4	69.2	37.2	69.6	73.8	69.4	70.1	48.8	25.7	19.9	25.0	71.3	71.9	70.7	73.5	73.0
Qwen3-32B	58.8	71.1	67.0	67.0	67.9	67.3	70.7	60.0	56.3	64.1	50.5	58.3	58.5	66.5	60.7	63.4	66.2	54.4	66.3	62.6	64.3	42.3	26.0	25.3	16.2	67.2	67.7	67.3	66.8	64.5
Qwen3-32B-Think	64.6	73.8	71.4	70.6	69.6	71.1	74.8	69.7	63.6	68.7	70.1	70.1	67.5	66.5	68.2	69.7	69.7	67.7	72.8	68.9	69.0	52.6	24.1	17.9	34.2	67.7	70.6	68.5	72.8	70.7
Qwen3-235B	66.0	75.0	71.3	72.3	73.5	71.6	74.3	66.3	65.0	67.0	66.5	66.5	66.5	68.9	69.7	69.7	68.5	66.5	70.2	69.4	71.4	54.3	27.9	38.3	45.1	72.6	70.6	70.6	71.9	71.4
Qwen3-235B-Think	**74.1**	**80.4**	**81.0**	**78.7**	79.3	**78.4**	**79.4**	**76.5**	**74.8**	**76.5**	**75.7**	**78.9**	**76.5**	**76.0**	**75.9**	**78.2**	71.9	78.9	**81.0**	**78.1**	**77.9**	68.9	37.1	48.0	44.7	**78.4**	**79.1**	79.3	**80.6**	**79.9**

Table 8: MMLU-ProX Lite 5-shot Results.

Model	Avg	EN	FR	DE	ES	PT	IT	HI	BN	UR	TE	MR	NE	ZH	JA	KO	VI	TH	ID	AR	AF	SW	WO	YO	ZU	RU	UK	SR	CS	HU
GPT-4.1	69.2	76.0	76.7	73.8	72.3	77.2	74.5	74.1	70.1	71.6	55.3	72.1	73.8	64.3	71.4	72.6	76.2	72.3	68.5	74.8	74.1	68.2	**44.2**	45.2	60.9	47.4	74.3	73.8	74.5	76.0
o4-mini	67.2	74.5	73.0	70.9	72.8	71.8	74.0	69.9	70.1	69.0	68.4	69.0	70.9	72.1	68.9	69.9	71.4	67.9	73.3	73.0	71.3	66.0	17.5	**54.6**	62.8	38.1	72.3	71.4	71.9	72.3
DeepSeek-R1	71.7	77.6	**80.8**	70.2	70.1	72.6	77.2	74.3	69.4	**75.5**	68.7	74.7	69.2	**75.2**	70.2	73.1	**78.1**	74.3	**80.6**	**75.2**	77.9	**71.6**	33.0	53.2	**66.8**	66.0	76.2	78.1	72.1	76.7
DeepSeek-V3	70.1	79.1	77.2	77.7	76.4	78.1	78.9	75.2	69.7	70.9	65.0	66.2	66.0	69.7	73.3	73.6	78.2	70.9	76.7	74.3	73.5	62.8	39.8	38.3	51.7	75.2	74.3	74.0	76.7	70.2
Phi4-14B	41.1	58.7	58.2	56.5	57.3	58.7	57.5	35.7	25.3	36.4	3.4	34.5	26.5	46.9	47.3	53.6	41.3	39.3	56.6	45.1	45.4	24.7	5.3	14.5	13.4	40.3	58.2	45.6	59.0	47.4
Gemma3-27B	56.4	65.8	63.1	60.5	63.8	62.8	62.1	59.0	56.3	58.5	57.8	57.8	57.0	54.1	59.4	58.5	61.9	54.1	61.9	56.1	61.4	53.2	4.1	33.8	44.0	63.8	60.5	61.6	62.1	59.9
Llama3.3-70B	52.9	66.3	67.7	63.6	62.9	65.6	66.7	56.5	44.2	48.6	12.8	55.6	43.5	60.2	54.4	56.0	61.6	58.8	63.6	59.4	65.0	53.9	23.3	32.3	31.8	28.1	62.6	60.2	63.9	46.3
Llama4-Scout	62.1	75.3	69.4	70.6	68.0	70.2	70.1	61.6	59.5	56.5	58.8	61.4	63.6	49.3	65.3	63.3	67.5	64.8	70.7	64.5	68.2	61.6	11.1	34.5	51.9	68.5	68.2	70.7	68.9	67.2
Qwen2.5-72B	58.9	73.1	68.0	63.9	68.4	68.7	69.0	60.0	58.0	59.5	38.6	53.9	55.8	67.3	62.8	65.3	67.7	62.1	66.8	64.8	67.5	42.0	20.1	25.7	27.2	69.9	65.8	67.0	67.7	61.2
Qwen3-30B-A3B	47.9	65.1	61.9	64.6	64.6	65.6	67.9	46.1	42.7	38.6	32.8	39.1	33.5	66.3	62.9	63.3	62.2	41.8	46.1	55.8	53.2	21.6	0.0	0.9	6.6	57.3	50.0	59.4	61.7	57.0
Qwen3-30B-A3B-Think	64.1	76.7	73.5	72.3	73.0	72.3	75.9	69.4	69.4	70.1	68.9	70.1	64.5	66.5	70.6	71.6	73.6	70.7	73.6	71.3	72.1	55.3	0.7	0.0	13.9	71.3	71.6	72.4	74.7	71.9
Qwen3-32B	50.4	62.6	46.4	51.5	69.7	67.2	65.8	54.9	42.0	51.0	27.7	51.4	54.8	65.5	60.4	64.8	67.5	41.8	53.9	62.9	54.6	28.9	2.4	8.3	24.7	26.7	57.1	66.0	68.9	61.4
Qwen3-32B-Think	60.4	70.9	69.0	66.5	69.4	67.3	71.4	65.8	63.9	66.3	53.6	63.8	62.2	65.6	67.9	70.6	65.3	63.1	70.9	67.2	66.7	56.0	0.7	7.5	20.2	65.1	70.6	68.4	69.6	65.1
Qwen3-235B	62.0	72.8	71.9	70.1	72.1	69.9	73.0	68.9	58.5	68.4	48.8	67.0	66.2	66.7	54.9	71.8	68.4	50.9	63.9	71.4	72.6	52.9	12.4	32.0	31.5	73.3	61.7	70.4	70.2	66.7
Qwen3-235B-Think	**73.4**	**80.1**	77.2	**78.7**	**80.4**	**80.1**	**83.0**	**77.2**	**75.7**	75.2	**77.9**	**76.0**	**77.6**	74.3	**75.7**	**77.0**	75.3	**77.9**	80.1	74.8	**78.2**	70.6	27.9	50.9	32.7	**77.4**	**78.1**	**81.0**	**80.3**	**77.7**

Table 9: MMLU-ProX Lite zero-shot Results.