MATH-IDN: A Multilingual Mathematical Problem Solving Dataset Featuring Local Languages in Indonesia

Xiao Xiao[♣] **Iftitahu Ni'mah**[♦♠] **Yuyun Wabula**[♠] **Mykola Pechenizkiy**[♦] **Meng Fang**[♣]

[♣] University of Liverpool [♦] Eindhoven University of Technology
[♠] Research Center for Data and Information Science, BRIN Indonesia
Xiao.Xiao@liverpool.ac.uk, i.nimah@tue.nl

Abstract

Large Language Models (LLMs) excel at mathematical reasoning in English, but their performance in low-resource languages remains underexplored. This gap is particularly critical in the Indonesian context, where equitable access to AI systems depends on robust multilingual reasoning across diverse local languages. We introduce MATH-IDN, a multilingual benchmark for mathematical problem solving in Indonesian, Javanese, Sundanese, and Buginese, with English as a reference, following the MATH dataset. We evaluate multiple open-source LLMs, including math-specialized, Southeast-Asian-adapted, and general-purpose models, under a zero-shot chain-of-thought setting. Results show that MATH-IDN presents a challenging and discriminative benchmark, revealing substantial performance gaps in low-resource languages, particularly Buginese, and highlighting key limitations in current multilingual reasoning capabilities. [1]

1 Introduction

Recent advances in mathematical reasoning within natural language processing, particularly research leveraging Large Language Models (LLMs) (Yuan et al., 2023; Touvron et al., 2023; Ouyang et al., 2022), have focused primarily on high-resource languages such as English. Prominent benchmarks including MATHQA (Amini et al., 2019), GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and MathOdyssey (Fang et al., 2025) have driven remarkable progress, yet they largely overlook the linguistic diversity of low-resource regions (Guo et al., 2024). To understand whether LLMs maintain cross-lingual consistency, that is, whether their reasoning proficiency in English degrades when applied to other languages there is a pressing need for multilingual benchmark datasets that enable systematic evaluation of reasoning across languages.

This need is particularly acute for Indonesia, one of the most linguistically diverse countries in the world. Although several new benchmarks have begun to explore multilingual mathematical reasoning, such as MMATH (Luo et al., 2025), MCLM (Son et al., 2025), MGSM (human-translated) (Shi et al., 2023), and mCOT-MATH (synthetic) (Lai and Nissim, 2024), none include the languages spoken in Indonesia. For Southeast Asia more broadly, resources such as SeaExam, SeaBench (Liu et al., 2025; Nguyen et al., 2024), and SeaCrowd (Lovenia et al., 2024) have expanded multilingual coverage but focus mainly on general NLP tasks rather than mathematical reasoning. Within Indonesia, NusaCrowd (Cahyawijaya et al., 2023) and IndoMMLU (Koto et al., 2023) represent major advances: NusaCrowd standardizes data across more than 700 local languages, while IndoMMLU benchmarks 64 academic subjects from the national curriculum. However, neither addresses symbolic or numerical reasoning, and both overlook the semantic and cultural shifts that arise in cross-lingual translation-leaving mathematical reasoning in Indonesian local languages a crucial yet unexplored challenge.

To address this gap, we introduce MATH-IDN, a multilingual mathematical reasoning dataset manually translated and human-verified from the MATH dataset (Hendrycks et al., 2021). The dataset covers five parallel languages, including English, Indonesian, Javanese, Sundanese, and Buginese, allowing systematic analysis of LLMs cross-lingual consistency in mathematical problem solving. Following the MATH500 dataset, MATH-IDN contains 500 expert-curated problems.

We conduct the zero-shot chain-of-thought evaluation on multiple LLMs, including LLaMA-3.1

[1] Our data and code are available at https://github.com/aialt/MATH-IND.

Findings of the Association for Computational Linguistics: EACL 2026, pages 4432–4438

Language	System prompt
English	You are a mathematics expert. Please solve the following problem and provide a detailed solution.
Indonesian	Kamu adalah ahli matematika. Tolong selesaikan persoalan matematika berikut dan berikan solusinya secara rinci.
Javanese	Awakmu ahli matematika. Tulung rampungna soal matematika ing ngisor iki lan wenehi solusine kanthi rinci.
Sundanese	Anjeun ahli matematika. Mangga tuntaskeun soal matematika ieu sarta sadiakeun solusi lengkepna
Bugis	Iko ahli matematika. Pappurai persoalang matematika e nappa alengka solusi na.

Table 1: System prompt used for each language's mathematical reasoning task.

(Dubey et al., 2024), Gemma-2 (Rivière et al., 2024), Qwen2.5 (Qwen, 2024), Qwen2.5-Math (Yang et al., 2024), and Southeast-Asian models such as Bakpia, Komodo (Owen et al., 2024), and SEA-LION (Ong and Limkonchotiwat, 2023). Our results reveal substantial performance degradation in low-resource languages, particularly Buginese, highlighting MATH-IDN as a challenging benchmark for cross-lingual reasoning research.

2 MATH-IDN Dataset

2.1 Task Definition

A mathematical reasoning task in Natural Language Processing (NLP) aims to develop a computational model that can solve mathematical problems presented in natural language (Hendrycks et al., 2021). Formally, given a mathematical problem expressed as a natural question or query Q_i, the task is to generate a valid answer A_i in a model's response. A complete response consists of two main components: (1) An intermediate reasoning process or Chain-of-Thought (CoT) R_i; and (2) A final answer A_i.

To study the consistency of LLMs mathematical reasoning across languages, we use multilingual system prompts tailored to each target language to instruct the models to solve the reasoning tasks, as shown in Table 1.

2.2 Dataset Construction

Our data collection process consists of several steps. First, we use MATH500 examples from the existing mathematical reasoning dataset MATH (Hendrycks et al., 2021). Second, we use gpt-4.1-2025-04-14 to translate samples from English to the target local languages in Indonesia. Third, given the translation results, we ask the native speakers of the target languages, as human experts, to analyze translation quality and refine the translated texts.

LLM Translation We use an instruction template, as shown in Table 2, to prompt gpt-4.1 with a one-shot example.

```
{% block instruction %}
{{instruction_start}}
Terjemahkan teks Bahasa Inggris berikut menjadi teks Bahasa Indonesia. Jangan mengubah teks berformat latex maupun ekspresi matematika di dalamnya.
{{instruction_end}}
{% endblock %}

{# example 1 #}
{{input_start}}
Convert the point $(0,3)$ in rectangular coordinates to polar coordinates. Enter your answer in the form $(r,\theta)$, where $r > 0$ and $0 \leq \theta < 2\pi$. Terjemahan:
{{input_end}}
{{output_start}}
Konversikan titik $(0,3)$ dalam koordinat persegi panjang ke koordinat polar. Masukkan jawaban Anda dalam bentuk $(r,\theta)$, di mana $r > 0$ dan $0 \leq \theta < 2\pi$.
{{output_end}}
```

Table 2: Example prompt template used for translating samples from English into Indonesian.

Annotator Recruitment Translating texts to under-represented languages, such as local languages in Indonesia, requires access to native speakers who can at least speak in two languages: English and a local language. Our annotators, including university and institutional researchers, have been working on low-resource translation tasks: one translator for Indonesian and Javanese; two translators for Sundanese; and two translators for Buginese. We use at least one hour of training and supervision to familiarize the translators with scoring system because the characteristic of the mathematical reasoning dataset differs from other translation datasets. The five-scaled scoring system is to measure the quality of LLM translation results before being refined by human translators.

Translation Quality Rubrics For the translation task, we use five-scaled scoring system to

measure the quality of LLM translation results. After that, the translations are refined by human translators. For each sample, the translators are asked to rate its "Fluency" and "Change of Meaning".

For "Fluency", the rubrics are as follows:

- 5 - Excellent: The text is perfectly natural and fluent, indistinguishable from human-written text. It is free of grammatical errors and awkward phrasing.
- 4 - Good: The text is fluent and easily understood. It may contain minor grammatical errors or typos that do not impact the overall meaning.
- 3 - Adequate: The text is comprehensible, but contains noticeable errors in grammar or unnatural word choices. The phrasing may be awkward, similar to that of a non-native speaker.
- 2 - Disfluent: The text contains significant grammatical errors that make it difficult to understand. The reader must exert considerable effort to grasp the intended meaning.
- 1 - Incomprehensible: The text is nonsensical or so grammatically flawed that its meaning cannot be understood.

For "Change of Meaning", the rubrics are as follows:

- 5 - No Change: The translation perfectly preserves the full semantic content and nuance of the original text.
- 4 - Mostly Preserved: The translation retains the core meaning and most details, but with minor, non-critical differences in nuance or emphasis.
- 3 - Partially Preserved: The core meaning is recognizable, but significant details are omitted or altered, leading to a partial loss of information.
- 2 - Misleading: The translation introduces inaccuracies or ambiguities that significantly distort the original meaning, potentially leading to misinterpretation.
- 1 - Completely Different: The translation conveys a meaning that is entirely different from, or contradictory to, the original text.

3 Experiments

3.1 Models

We select a range of open-source LLMs across three categories: general-purpose models, including LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Gemma-2-9B-IT (Rivière et al., 2024), and Qwen2.5-7B-Instruct (Qwen, 2024); math-specific models, represented by Qwen2.5-Math-7B-Instruct (Yang et al., 2024); and SEA-specific models, including Bakpia-V1-1.5B-Javanese, Komodo-7B-Base (Owen et al., 2024), and Llama-SEA-LION-v3.5-8B-R (Ong and Limkonchotiwat, 2023).

3.2 Metrics

We evaluate multilingual mathematical problem solving along two dimensions: (i) *language-wise correctness* on each target language, and (ii) *cross-lingual stability* of correctness across the parallel language set. Let $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ be the $N = 500$ problems in MATH-IDN, and let $\mathcal{L} = \{\text{EN}, \text{ID}, \text{JV}, \text{SU}, \text{BG}\}$ denote English, Indonesian, Javanese, Sundanese, and Buginese. For each model and each $(q_i, \ell) \in \mathcal{Q} \times \mathcal{L}$, we obtain a response and extract a final predicted answer $\hat{a}_{i,\ell}$ using the same answer parsing protocol as the MATH benchmark. We define an indicator of correctness

$$\mathbf{c}_{i,\ell} = \mathbb{I}(\hat{a}_{i,\ell} = a_i), \tag{1}$$

where a_i is the gold answer.

Accuracy. For each language $\ell \in \mathcal{L}$, we report the standard accuracy:

$$\text{Acc}(\ell) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{c}_{i,\ell}. \tag{2}$$

We also report per-language deltas relative to English, $\Delta(\ell) = \text{Acc}(\ell) - \text{Acc}(\text{EN})$.

Consistency (Cross-lingual correctness stability). We quantify whether a model answers the *same underlying problem* correctly across all parallel languages. We define the *Consistency Correct* rate as the proportion of problems answered correctly in *every* evaluated language:

$$\text{CC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\Big(\bigwedge_{\ell \in \mathcal{L}} \mathbf{c}_{i,\ell} = 1 \Big). \tag{3}$$

We define the complementary *Consistency Incorrect* rate as the proportion of problems for which correctness is *not* consistent across languages (i.e., correct in some languages but incorrect in others):

$$\text{CI} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\Big(\exists\, \ell, \ell' \in \mathcal{L} \text{ s.t. } \mathbf{c}_{i,\ell} \neq \mathbf{c}_{i,\ell'} \Big). \tag{4}$$

Intuitively, CC measures cross-lingual robustness, while CI captures language sensitivity.

Unique Correct (language-specific success). To localize which language uniquely enables success, we report *Unique Correct* counts and rates. A problem is *uniquely correct* in language ℓ if it is correct in ℓ and incorrect in all other languages:

$$\mathrm{UC}(\ell) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big(\mathbf{c}_{i,\ell} = 1 \wedge \bigwedge_{\ell' \in \mathcal{L}\setminus\{\ell\}} \mathbf{c}_{i,\ell'} = 0\Big). \tag{5}$$

Pivot Accuracy (decoupling language understanding from reasoning). To probe whether errors primarily arise from multilingual understanding rather than mathematical reasoning, we perform a pivot-language evaluation for source languages $\ell \in \{\mathrm{JV}, \mathrm{SU}, \mathrm{BG}\}$. We translate the source problem text into a pivot language $p \in \{\mathrm{EN}, \mathrm{ID}\}$ using gpt-4.1, then evaluate the target LLM on the pivoted input and compute accuracy:

$$\mathrm{PivotAcc}(\ell \to p) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{a}_i^{(\ell\to p)} = a_i), \tag{6}$$

where $\hat{a}_i^{(\ell\to p)}$ is the model prediction given the pivot-translated input.

3.3 Results

Overall Multilingual Accuracy Table 3 reports accuracy across English, Indonesian, and three local Indonesian languages. Across all evaluated models, performance consistently degrades when moving from English to low-resource languages, with the largest drops observed in Buginese. This confirms that MATH-IDN constitutes a challenging benchmark for multilingual mathematical reasoning.

While some models such as Gemma-2-9b-it and Qwen2.5-7B-Instruct exhibit comparable or slightly improved accuracy in Indonesian relative to English, all models experience non-trivial degradation in at least one local language. This pattern suggests that multilingual robustness does not transfer uniformly across linguistically related languages.

Domain specialization and regional alignment both contribute to improved multilingual performance. Math-specialized models such as Qwen2.5-Math-7B-Instruct achieve the highest overall accuracy in English and maintain smaller performance gaps across languages compared to their general-purpose counterparts. Similarly, SEA-adapted models such as Llama-SEA-LION-v3.5-8B-R exhibit stronger performance in Indonesian and Javanese than general-purpose models of similar scale. regional language exposure alone is insufficient for complex mathematical reasoning.

Cross-Lingual Consistency Analysis Accuracy alone does not capture whether models solve the same problems consistently across languages. Table 4 therefore reports cross-lingual consistency metrics. Qwen2.5-Math-7B-Instruct achieves the highest Consistency Correct (CC) rate (40.80%), indicating that domain-aligned training substantially improves cross-lingual stability. Similarly, Qwen2.5-7B-Instruct and Gemma-2-9b-it exhibit relatively high CC values, suggesting that larger-scale pretraining combined with instruction tuning contributes to more consistent multilingual behavior. In contrast, SEA-specific models such as Bakpia-V1-1.5B-Javanese and Komodo-7B-Base achieve markedly lower CC rates (below 20%), reflecting limited ability to generalize reasoning knowledge across languages.

High Consistency Incorrect (CI) rates are observed across most models, particularly general-purpose LLMs such as LLaMA-3.1-8B-Instruct. This indicates that even when models possess the underlying mathematical knowledge, their success is often conditional on the language of presentation. Notably, CI remains substantial even for models with strong English performance, highlighting that high monolingual accuracy does not guarantee multilingual robustness.

To further localize language-dependent effects, we examine Unique Correct (UC) cases, where a problem is answered correctly in exactly one language and incorrectly in all others. Across all models, English accounts for the majority of UC instances, whereas low-resource languagesespecially Bugineserarely produce unique successes. This asymmetry suggests that mathematical reasoning knowledge in current LLMs is disproportionately anchored in high-resource languages, with limited independent grounding in local languages. SEA-adapted models reduce this imbalance to some extent in Indonesian and Javanese, but the effect does not extend reliably to more under-represented languages.

Pivot language performance analysis To disentangle mathematical reasoning ability from mul-

	Models	English	Indonesian	Javanese	Sundanese	Buginese
General purpose	LLaMA-3.1-8B-Instruct	60.80%	48.40% (−12.40%)	42.80% (−18.00%)	37.60% (−23.20%)	39.00% (−21.80%)
	Gemma-2-9b-it	53.40%	58.20% (+4.80%)	54.60% (+1.20%)	56.80% (+3.40%)	47.60% (−5.80%)
	Qwen2.5-7B-Instruct	60.80%	62.40% (+1.60%)	55.80% (−5.00%)	56.40% (−4.40%)	49.00% (−11.80%)
Math-Specific	Qwen2.5-Math-7B-Instruct	64.60%	61.80% (−2.80%)	55.60% (−9.00%)	58.80% (−5.60%)	50.20% (−14.40%)
SEA-specific	Bakpia-V1-1.5B-Javanese	39.80%	30.20% (−9.60%)	27.40% (−12.40%)	26.80% (−13.00%)	24.00% (−15.80%)
	Komodo-7B-Base	19.80%	20.60% (+0.80%)	19.40% (−0.40%)	18.40% (−1.40%)	17.00% (−2.80%)
	Llama-SEA-LION-v3.5-8B-R	61.20%	57.80% (−3.40%)	54.80% (−6.40%)	48.60% (−12.60%)	42.00% (−19.20%)

Table 3: Performance of LLMs across languages under zero-shot chain-of-thought evaluation. Numbers denote accuracy (%), with absolute differences from English accuracy shown in parentheses ($\Delta(\ell)$).

Model	**Consistency Correct (CC)**	**Consistency Incorrect (CI)**	**Unique Correct (UC)**				
			English	**Indonesian**	**Javanese**	**Sundanese**	**Buginese**
LLaMA-3.1-8B-Instruct	23.60%	47.80%	47 / 9.40%	8 / 1.60%	11 / 2.20%	2 / 0.40%	9 / 1.80%
Gemma-2-9b-it	33.40%	37.40%	10 / 2.00%	10 / 2.00%	2 / 0.40%	5 / 1.00%	5 / 1.00%
Qwen2.5-7B-Instruct	35.60%	38.60%	12 / 2.40%	11 / 2.20%	3 / 0.60%	4 / 0.80%	5 / 1.00%
Qwen2.5-Math-7B-Instruct	40.80%	33.80%	13 / 2.60%	6 / 1.20%	5 / 1.00%	7 / 1.40%	7 / 1.40%
Bakpia-V1-1.5B-Javanese	18.40%	29.40%	46 / 9.20%	12 / 2.40%	8 / 1.60%	6 / 1.20%	1 / 0.20%
Komodo-7B-Base	15.20%	12.60%	13 / 2.60%	13 / 2.60%	10 / 2.00%	5 / 1.00%	1 / 0.20%
Llama-SEA-LION-v3.5-8B-R	28.20%	45.00%	17 / 3.40%	9 / 1.80%	8 / 1.60%	7 / 1.40%	1 / 0.20%

Table 4: Cross-lingual consistency analysis across five languages.

Language	Models	Model Name	Acc - Original	PivotAcc (EN)	PivotAcc (IDN)
Javanese	general	Gemma-2-9b-it	54.60%	56.00%	56.60%
Sundanese	general	Gemma-2-9b-it	56.80%	57.20%	57.60%
Buginese	general	Gemma-2-9b-it	47.60%	52.00%	53.00%
Javanese	math-specific	Qwen2.5-Math-7B-Instruct	55.60%	64.60%	61.20%
Sundanese	math-specific	Qwen2.5-Math-7B-Instruct	58.80%	62.20%	61.80%
Buginese	math-specific	Qwen2.5-Math-7B-Instruct	50.20%	58.60%	58.40%
Javanese	SEA-specific	Llama-SEA-LION-v3.5-8B-R	54.80%	64.20%	58.60%
Sundanese	SEA-specific	Llama-SEA-LION-v3.5-8B-R	48.60%	63.00%	58.20%
Buginese	SEA-specific	Llama-SEA-LION-v3.5-8B-R	42.00%	58.20%	55.00%

Table 5: Pivot-language evaluation results. Acc - Original denotes accuracy on the original low-resource language input. Pivot Acc (EN/IDN) denotes accuracy after translating the input into English or Indonesian before inference, isolating multilingual comprehension effects from reasoning ability.

tilingual comprehension, we conduct a pivot-language evaluation, as shown in Table 5. Across all evaluated models and languages, pivoting low-resource inputs through English or Indonesian consistently improves accuracy, often substantially. For example, Qwen2.5-Math-7B-Instruct shows accuracy gains of more than 8% in Buginese when using English as a pivot language. This pattern provides strong evidence that multilingual understanding, rather than mathematical reasoning itself, constitutes the primary bottleneck for low-resource performance. Together with the consistency analysis, these results demonstrate that many errors arise from lexical and semantic interpretation failures rather than deficiencies in logical reasoning.

4 Conclusion

We introduced MATH-IDN, a multilingual benchmark for evaluating mathematical problem solving in Indonesian and under-represented local languages. Built on expert-verified translations of MATH, MATH-IDN enables systematic analysis of both accuracy and cross-lingual consistency under controlled conditions. Our results show that current LLMs suffer substantial performance degradation in low-resource languages and frequently exhibit language-dependent correctness. Consistency and unique-correct analyses reveal that reasoning success is often anchored in high-resource languages, while pivot-language evaluation provides evidence that multilingual comprehension, rather than mathematical reasoning, is the dominant bottleneck. We hope MATH-IDN will support future work on more robust and equitable multilingual reasoning.

5 Limitations

The current evaluation is restricted to language models with fewer than 10 billion parameters. This limitation primarily arises from computational and resource constraints during experimentation. In addition, there are 700+ languages in Indonesia. However, our work has only focused on a small fraction of these languages. In addition, there are other regional languages in Indonesia, such as Balinese used in Bali and Minangkabau spoken in West Sumatra.

6 Acknowledgement

This research has been supported by a collaboration between Research Center for Data and Information Science (BRIN Indonesia) and PT. Datasaur Software Indonesia under the Cooperation Agreement No. 598/V/KS/2025. PT. Datasaur Software Indonesia provided in-kind support in the form of an annotation tool used by the authors for semi-automated data labeling.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, and 29 others. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2025. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *Scientific Data*, 12(1):1392.

Jia Guo, Longxu Dou, Guangtao Zeng, Stanley Kok, Wei Lu, and Qian Liu. 2024. Sailcompass: Towards reproducible and robust evaluation for southeast asian languages. *arXiv preprint arXiv:2412.01186*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

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.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6119–6136, Albuquerque, New Mexico. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2025. Mmath: A multilin-

gual benchmark for mathematical reasoning. *arXiv preprint arXiv:2505.19126*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

David Ong and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into indonesia's regional languages. *arXiv preprint arXiv:2403.09362*.

Qwen. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

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. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14333–14368. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.