MalayMMLU: A Multitask Benchmark for the Low-Resource *Malay* Language

Anonymous ACL submission

Abstract

Large Language Models (LLMs) exhibit advanced proficiency in language reasoning and comprehension across a wide array of languages. While their performance is notably robust in well-resourced languages, the capabilities of LLMs in low-resource languages, such as Bahasa Malaysia (hereinafter referred to as *Malay*), remain less explored due to a scarcity of dedicated studies and benchmarks. To enhance our understanding of LLMs' performance in Malay, we introduce the first multitask language understanding benchmark specifically for this language, named MalayMMLU. This benchmark comprises 24,213 questions spanning both primary (Year 1-6) and secondary (Form 1-5) education levels in Malaysia, encompassing 5 broad topics that further divide into 22 subjects. We conducted an empirical evaluation of 18 LLMs, assessing their proficiency in both Malay and the nuanced contexts of Malaysian culture using this benchmark. We will release the MalayMMLU benchmark and the corresponding code publicly upon paper acceptance.

1 Introduction

Large Language Models (LLMs) like GPT-4 (OpenAI et al., 2024), LLaMA (Touvron et al., 2023), and Mistral (Jiang et al., 2023) are renowned for their proficiency in various benchmarks related to language understanding (Wang et al., 2018; Hendrycks et al., 2021) and question answering (Rajpurkar et al., 2018; Talmor et al., 2019). These models excel in fields such as science, humanities, business, and mathematics due to their training on multilingual datasets predominantly comprising well-resourced languages like English and Chinese. However, their performance in low-resource languages, such as Bahasa Malaysia (hereafter referred to as Malay), which is widely used in Malaysia, has been inadequate (see Table 4).

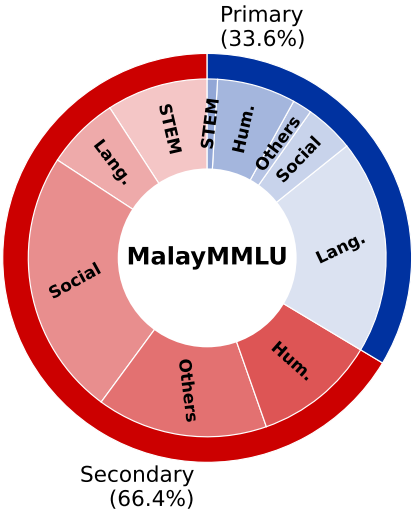

Figure 1: Data distribution by education level and topics in MalayMMLU benchmark. MalayMMLU contains 22 subjects that are categorized into topics such as Language (Lang.), Humanities (Hum.), STEM, Social Science (Social) and Others.

Despite ongoing research into multilingual LLMs, there remains a significant gap in systematic and comprehensive benchmarks for low-resource languages comparable to the Multitask Machine Learning Understanding (MMLU) framework. This gap impedes the evaluation of LLMs' reasoning capabilities in these languages.

For instance, the SeaLLMs initiative (Nguyen et al., 2023) is designed to boost the multilingual capabilities of LLMs across Southeast Asia, focusing on languages such as Indonesian, Thai, Vietnamese, English, and Chinese. However, the initiative's training corpus comprises less than 2% Malay content, significantly ten times less than that for Indonesian. Furthermore, its evaluation platform, SeaBench, contains fewer than 100 Malay language questions, suggesting that the initiative may not provide a comprehensive assessment of Malay language capabilities.

Similarly, the IndoMMLU project (Koto et al., 2023) has advanced the evaluation of LLMs in In-

Education Level	Topic	Count
Primary	Language	4684
	Humanities	1721
	Social science	1078
	Others	426
	STEM	224
Secondary	Social science	5840
	Others	3743
	Humanities	2674
	STEM	2219
	Language	1604
Total		24,213

Table 1: Data distribution by education level and topics in MalayMMLU benchmark.

donesian and other regional languages, including Madurese, Makassarese, and Balinese. This comprehensive evaluation has demonstrated that even sophisticated models like GPT-3.5 encounter difficulties with high school-level examinations in these specific linguistic and cultural contexts, emphasizing the substantial challenges LLMs face in adapting to local nuances.

Given that Malay is the official language of Malaysia and is spoken by over 30 million people, it is crucial yet underexplored in linguistic research. Prior initiatives, including SeaLLMs and Sailor (Dou et al., 2024), have attempted to integrate Malay into their datasets, but the proportion of Malay data remains below 5%.

To address this research deficiency, we introduce MalayMMLU, a benchmark consisting of 24,213 multiple-choice questions from primary to secondary education levels in Malaysia, covering five topics subdivided into 22 subjects. This benchmark aims to rigorously assess the proficiency of LLMs in Malay language (please refer to Figure 1 and Table 1).

Our contributions are as follows:

- We introduce MalayMMLU, the first dedicated benchmark for the Malay language, featuring 24,213 questions across five topics and 22 subjects at different educational levels. This novel benchmark enables detailed assessments of language understanding in Malay.
- Our empirical evaluation of 18 LLMs highlights GPT-4 outperforms others by approximately 13% and shows the advantages of regional dataset training (refer Table 4).
- We analyze how question length, number of options, and educational levels impact LLM performance, noting a decline in accuracy as these factors increase. This provides insights into LLM scalability and task complexity handling (refer Section 5.2.)
- By comparing LLMs on Malay and Indonesian (two closely related languages[1]), we examine the effects of lexical similarities and cultural nuances on model effectiveness, enriching our understanding of language model training across very similar languages (refer Table 7).

2 Related works

Evaluation benchmarks. LLMs are acclaimed for their human-like proficiency in language understanding and reasoning (OpenAI et al., 2024; Touvron et al., 2023; Jiang et al., 2023). As these models advance, systematic evaluations of their linguistic capabilities are increasingly essential. Benchmarks such as GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016) have traditionally assessed language models' (LMs) abilities in natural language understanding (NLU) and question answering (QA), respectively.

With the continuous improvement of LMs, models have excelled in these benchmarks, creating a demand for more challenging and comprehensive evaluations. XGLUE (Liang et al., 2020) and XTREME-R (Ruder et al., 2021) introduced multilingual benchmarks to evaluate LMs' cross-lingual capabilities. While these benchmarks are invaluable for assessing language performance across languages, they do not thoroughly test LMs on broader aspects such as world knowledge, commonsense reasoning, mathematics, and coding. Recent benchmarks like MMLU (Hendrycks et al., 2021), CommonsenseQA (Talmor et al., 2019), TriviaQA (Joshi et al., 2017), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021) provide more comprehensive evaluations across these various domains. However, these evaluations are predominantly in English, leading to a gap in understanding LLMs' capabilities in other languages. For example, IndoMMLU (Koto et al., 2023) revealed that while LLMs perform adequately on English-based MMLU (Hendrycks et al., 2021), their performance significantly declines when assessed in Indonesian.

[1]Malay and Indonesian are mutually intelligible, with differences mainly in vocabulary, pronunciation and spelling. Please check http://alturl.com/2wfh9 for more details.

Low-Resource Languages. Low-resource languages, characterized by a scarcity of available datasets, pose unique challenges for LLM development. English dominates online content, comprising about 50% of web content[2]. In contrast, Southeast Asian languages such as Indonesian and Vietnamese represent only around 1% of web content. Malay, even less prevalent, accounts for a mere 0.1%, ten times less than Indonesian.

Although initiatives like SeaLLMs (Nguyen et al., 2023) and Sailor (Dou et al., 2024) have made strides in incorporating Malay into their pretraining datasets, these efforts are limited, with only about 1% and 4% Malay content, respectively. Consequently, the evaluations of LLMs in Malay are constrained, and comprehensive linguistic datasets in Malay are extremely scarce. This paucity hinders a thorough assessment of LLMs' performance in the Malay language.

Language Similarity. Malay and Indonesian share a high degree of lexical similarity, approximately 90% (Omar, 2001). Studies by Ranaivo-Malancon and Lin et al. highlighted the existence of numerous identical words with differing meanings in both languages. Despite these similarities, the impact on LLM performance remains largely unexplored. Understanding how these linguistic similarities affect LLMs' handling of low-resource languages like Malay and Indonesian is crucial, yet remains an under-investigated area of research.

As a summary, these insights underscore the critical necessity of establishing comprehensive benchmarks like MalayMMLU to rigorously evaluate LLMs in low-resource languages.

3 Bahasa Malaysia: National Language Context and Usage Overview

Malay, the national language of Malaysia, remains significantly underexplored in computational linguistics and natural language processing research. Known as *Bahasa Malaysia* in official contexts, Malay serves as the primary medium for government announcements, documents, and official communications across Malaysia. This extensive usage underscores its central role in Malaysian public life and governance.

In the educational system, Malay is a mandatory subject from primary through secondary school. The Malaysian education system mandates proficiency in Malay, requiring students to pass Malay language examinations to progress to tertiary education levels[3]. This requirement reflects Malay's crucial role in academic and professional advancement within Malaysia.

Furthermore, the *Bahasa Malaysia* curriculum encompasses a wide range of subjects, ensuring that students gain a deep and comprehensive understanding of the language. According to the Ministry of Education Malaysia[4], the curriculum is designed not only to promote linguistic proficiency but also to instill a deep appreciation for Malay literature, culture, and heritage. The language's prominence extends to various national examinations, including the *Sijil Pelajaran Malaysia (SPM)* and *Pentaksiran Tingkatan 3 (PT3)*, which are critical milestones for Malaysian students.

Malay's status as a national language also translates into its usage in legal documents, media, and public signage, reinforcing its pervasive influence in everyday life. Despite its wide use and cultural significance, Malay has received limited attention in the development and evaluation of LLMs. As such, there is a pressing need for more dedicated research and resources to enhance the capabilities of LLMs in understanding and processing Malay, particularly in low-resource contexts.

4 MalayMMLU

Motivated by the scarcity of datasets in Malay, we propose MalayMMLU, a benchmark that comprises Malay-language questions contextualized for Malaysia, covering various education levels and subjects. Following the format of the English MMLU, we curated this dataset in alignment with the local educational curriculum.

The Malaysian curriculum is divided into two phases: (i) primary school level and (ii) secondary school level. The primary school level spans ages 7 to 12, while the secondary school level covers ages 13 to 17. For each level, we prepared the dataset in accordance with the standard curriculum set by the Ministry of Education, Malaysia[5].

By aligning the MalayMMLU with educational standards, we aim to establish a comprehensive benchmark for assessing LLMs' capabilities in understanding and processing the Malay language across various educational levels. This thorough

[2] http://alturl.com/tcwg4

[3] https://blog.mytutor.my/halatuju-pendidikan-spm-vs-igcse

[4] Website: https://www.moe.gov.my/

[5] Links to the curriculum: Primary school level and Secondary school level

Mathematics (Form 4)

Diberi set M = {2,3,4,5,6,7,8,9,10}. Satu nombor dipilih secara rawak daripada set itu.Cari kebarangkalian bahawa nombor yang terpilih itu ialah faktor bagi 32 **A. 1/3** B. 2/3 C. 2/9 D. 4/9	Given a set M = {2,3,4,5,6,7,8,9,10}. A number is chosen at random from the set. Find the probability that the chosen number is a factor of 32 **A. 1/3** B. 2/3 C. 2/9 D. 4/9

Chemistry (Form 4)

Larutan akueus sesuatu elektrolit mengandungi: * Anion dan kation elektrolit. * Ion hidrogen dan ion hidroksida daripada penceraian molekul air. Hanya satu kation dan satu anion yang akan dipilih untuk dinyahcas pada setiap elektrod. Antara faktor yang berikut,yang manakah mempengaruhi pemilihan ion untuk dinyahcas? I Kedudukan ion dalam siri elektrokimia. II Kepekatan ion di dalam elektrolit. III Isipadu elektrolit dalam sel elektrolisis. IV Kuantiti arus yang mengalir melalui elektrod. **A. I dan II sahaja** B. I dan IV sahaja C. II dan III sahaja D. II dan IV sahaja	An aqueous solution of an electrolyte contains: * Electrolyte anions and cations. * Hydrogen ions and hydroxide ions from the dissociation of water molecules. Only one cation and one anion will be selected to be discharged at each electrode. Which of the following factors affects the selection of ions to be discharged? I The position of ions in the electrochemical series. II Concentration of ions in the electrolyte. III The volume of the electrolyte in the electrolysis cell. IV The quantity of current flowing through the electrodes. **A. I and II only** B. I and IV only C. II and III only D. II and IV only

Figure 2: Sample questions in *Malay* (left) and their *English* translation (right). The correct answer is bolded.

evaluation is designed to contextualize LLM performance within the Malaysian educational framework, systematically testing these models against locally relevant curriculum and exam-style questions. Additionally, this benchmark enables researchers to pinpoint specific weaknesses of LLMs in the Malaysian context, underscoring the importance of developing models that are attuned to local nuances to better serve the Malaysian community. This targeted approach not only enhances model accuracy but also fosters LLMs that are more culturally and contextually relevant.

4.1 Data Preparation

We collected the dataset through an online learning platform widely adopted by most primary and secondary schools in Malaysia. On this platform, teachers can voluntarily upload practice exam questions they have created, along with the corresponding answers, and specify the education level.

The platform allows for various modes of questions, enabling teachers to include images, videos, and audio references. However, for the purpose of our benchmark, which focuses on unimodal, text-based evaluation, we excluded all questions containing images, videos, and audio. This ensures that our dataset remains consistent and suitable for text-based analysis.

Category	Subjects
STEM	Computer Science (Secondary), Biology (Secondary), Chemistry (Secondary), Computer Literacy (Secondary), Mathematics (Primary, Secondary), Additional Mathematics (Secondary), Design and Technology (Primary, Secondary), Core Science (Primary, Secondary), Information and Communication Technology (Primary), Automotive Technology (Secondary)
Language	Malay Language (Primary, Secondary)
Social science	Geography (Secondary), Local Studies (Primary), History (Primary, Secondary)
Others	Life Skills (Primary, Secondary), Principles of Accounting (Secondary), Economics (Secondary), Business (Secondary), Agriculture (Secondary)
Humanities	Quran and Sunnah (Secondary), Islam (Primary, Secondary), Sports Science Knowledge (Secondary)

Table 2: Fine-grained subjects by Category and Level. All subjects are labeled according to their respective education levels.

4.2 Data Cleaning and Standardization

To ensure our dataset quality, we implemented a data cleaning pipeline designed to standardize the dataset. The pipeline is designed as follows:

- Discard all questions with non-text contents such as images, videos, and audio.
- Exclude questions containing non-Latin characters, such as Arabic and Jawi, to focus on Malay content.
- Remove questions that do not provide options and corresponding answers.
- Filter out questions with external URLs.
- Strip HTML tags and irrelevant symbolic characters from the text.
- For questions lacking alphabetical options, generate them as necessary.
- Apply a deduplication algorithm using string matching to eliminate redundant questions, identifying and removing those with similarity above 85%.

After implementing the aforementioned pipeline, we conducted random sampling and manual verification of the processed questions. This process yielded a total of 24,213 questions for MalayMMLU spanning 22 subjects. Subsequently,

we categorized these subjects according to the predefined topics in MMLU (refer Figure 1).

4.3 Data Distribution

We first visualize the distribution of MalayMMLU according to the subjects and education levels, organized according to the MMLU format, as shown in Figure 1. We then present the exact count of each subject in Table 1. The dataset encompasses categories such as "Humanities", "Social Science", "Science, Technology, Engineering, and Mathematics" (STEM), "Others", and an additional category for "Language". Each category is further subdivided into detailed subjects, as depicted in Table 2, and their detailed descriptions are provided in Table 8 (see Appendix). We also depict sample questions and their corresponding English translations in Figure 2, where the correct answers are bolded.

Question length. In Table 3, we present the average length of questions across various topics and education levels. The data reveal a trend of increasing question length as educational levels progress, implying an enhancement in students' language comprehension with higher educational attainment. This suggests a correlation between the complexity of language use and the educational level.

Fine-grained subjects. In Table 8 (see Appendix), we illustrate the detailed distribution of subject-specific data. Each subject encompasses a minimum of 96 questions, providing a robust dataset to thoroughly assess the performance of LLMs within the context of Malaysia's standardized curriculum at both primary and secondary educational levels.

Group	Question	Answer
Primary school	107.69	13.71
Secondary school	144.73	18.37
STEM	142.78	17.55
Social science	150.78	19.01
Humanities	106.48	15.11
Language	116.47	13.64
Other	146.54	19.28

Table 3: Average question and answer length (in characters) for each education group and subject area. We observe the secondary school level has a longer question and answer length compared to the primary school level.

5 Experiments

5.1 Experimental Setup

We conduct a comprehensive study across current state-of-the-art models[6], under both zero-shot and few-shot settings. We study a total of *18* LLMs, including both *open-source* and *close-sourced* models. For **open-source** models, we include LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), SeaLLMs (Nguyen et al., 2023), Sailor (Dou et al., 2024), Phi (Abdin et al., 2024), Qwen (Bai et al., 2023), Gemma (Team et al., 2024), Komodo (Owen et al., 2024) and MaLLaM (Zolkepli et al., 2024); meanwhile for **close-sourced** models, we study both GPT-3.5, GPT-4 (OpenAI et al., 2024). For GPT-3.5 and GPT-4, we utilize `gpt-3.5-turbo-0125` and `gpt-4-turbo-2024-04-09` respectively. Among these models, SeaLLMs and Sailor are finetuned with multiple SEA languages dataset, while Komodo is finetuned solely on Indonesian languages and MaLLaM is finetuned on Malaysian languages which includes Malay, Chinese, English and Tamil. We include the artifacts of the evaluated models in Table 17 (Appendix).

Accuracy. For open-source models, we calculate their first token and full answer accuracy, following the implementation of IndoMMLU. For closed-source models, we employ string matching to calculate its first token and full answer accuracy.

Prompt. For MalayMMLU, we employ the prompt template: *"Berikut adalah soalan aneka pilihan tentang [SUBJECT]. Sila berikan jawapan sahaja."*, followed by the question and options. Our prompt template translates into *"The following is a multiple choice question for [SUBJECT]. Please provide the answer only."* For IndoMMLU, we reuse their prompt template.

5.2 Results

We report the zero-shot results of 18 LLMs on MalayMMLU, as depicted in Table 4. We calculate their first token accuracy, according to the topics, regardless of the education levels. The full answer accuracy is included in Table 9 (see Appendix).

Best performer. From Table 4, it is evident that GPT-4 achieved the highest first token accuracy, establishing it as the leading LLM for the Malay language. Among the open-source LLMs, Sailor-7B recorded the highest average scores, surpassing LLaMA-3-8B. This indicates that Sailor-7B, despite having a smaller model size compared to some peers, effectively captures and processes the linguistic features essential for understanding and

[6] As of June 2024.

Model	Language Acc.	Humanities Acc.	STEM Acc.	Social Science Acc.	Others Acc.	Average Acc.
GPT-4	**82.90**	**83.91**	**78.80**	**77.29**	**77.33**	**80.11**
GPT-3.5	69.62	71.01	67.17	66.70	63.73	67.78
LLaMA-3 (8B)	63.93	66.21	62.26	62.97	61.38	63.46
LLaMA-2 (13B)	45.58	50.72	44.13	44.55	40.87	45.26
LLaMA-2 (7B)	47.47	52.74	48.71	50.72	48.19	49.61
Mistral-v0.3 (7B)	56.97	59.29	57.14	58.28	56.56	57.71
Mistral-v0.2 (7B)	56.23	59.86	57.10	56.65	55.22	56.92
Sailor† (7B)	74.54	68.62	62.79	64.69	63.61	67.58
SeaLLM-v2.5† (7B)	69.75	67.94	65.29	62.66	63.61	65.89
Phi-3 (14B)	60.07	58.89	60.91	58.73	55.24	58.72
Phi-3 (3.8B)	52.24	55.52	54.81	53.70	51.74	53.43
Qwen-1.5 (7B)	60.13	59.14	58.62	54.26	54.67	57.18
Qwen-1.5 (4B)	48.39	52.01	51.37	50.00	49.10	49.93
Qwen-1.5 (1.8B)	42.70	43.37	43.68	43.12	44.42	43.34
Gemma (7B)	45.53	50.92	46.13	47.33	46.27	47.21
Gemma (2B)	46.50	51.15	49.20	48.06	48.79	48.46
Komodo† (7B)	43.62	45.53	39.34	39.75	39.48	41.72
MaLLaM-v2† (5B)	42.56	46.42	42.16	40.81	38.81	42.07

Table 4: Zero-shot results of various LLMs on MalayMMLU. We report the first token accuracies of the LLMs. Highest scores are **bolded** and second highest scores are underlined. † denotes LLMs finetuned with SEA datasets.

generating Malay language content.

LLMs finetuned with SEA datasets. Our analysis reveals that LLMs finetuned with Southeast Asian (SEA) datasets, such as Sailor and SeaLLMs exhibit enhanced performance in Language subjects, which coheres with the findings of (Koto et al., 2023). However, their performance in other topics is comparable to that of LLaMA-3-8B, which has not been finetuned with SEA datasets. This suggests that regional finetuning primarily boosts language processing capabilities, possibly due to better handling of regional linguistic nuances.

Additionally, our observations indicate that Komodo, which is finetuned exclusively on an Indonesian dataset, and MaLLaM, finetuned on a diverse dataset including Malay, Chinese, English, and Tamil, underperforms on the MalayMMLU dataset. This highlights potential areas for improvement, particularly in optimizing these models for broader linguistic adaptability and comprehension. The discrepancy in performance could stem from insufficient representation of Malay linguistic features in training datasets, suggesting the need for more balanced and comprehensive data inclusions.

Accuracies across Education Levels. In Figure 3, we present the performance of various LLMs segmented by educational levels, where levels 1-6 correspond to primary school (Year 1-6) and levels 7-11 pertain to secondary school (Form 1-5). We observe a notable decline in the accuracies of LLMs as the educational level increases from Year 1 to Form 5. This suggests an increase in the complexity and difficulty of questions at higher educational levels.

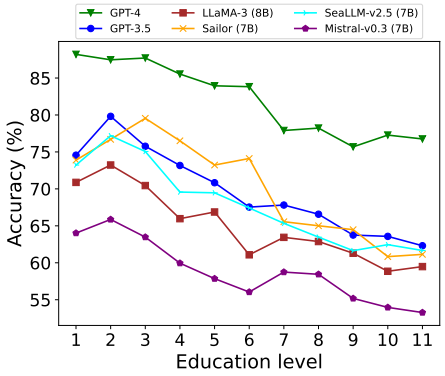

Figure 3: Accuracy of LLMs on MalayMMLU across different education level. Level 1-6 refer to primary school level (Year 1-6), while level 7-11 refer to secondary school level (Form 1-5). The education of 1 to 6 belong to primary school and level 7 to 11 belong to secondary school.

We hypothesize that this decrease in accuracy is indicative of the heightened cognitive and linguistic demands of questions designed for higher-level students, which may challenge the current capabilities of LLMs. These findings underscore the need for targeted improvements in model training, particularly in enhancing comprehension and pro-

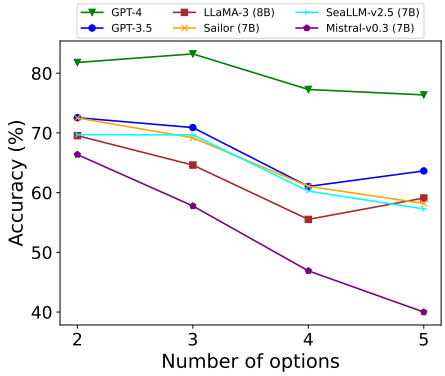

Figure 4: Accuracy of LLMs across different number of options. We observe that LLMs' performances generally decrease as the number of options increases.

Model	Correlation
GPT-4	-0.3331
GPT-3.5	-0.5339
LLaMA-3 (8B)	-0.5776
Sailor (7B)	-0.4813
SeaLLM-v2.5 (7B)	-0.4842
Mistral-v0.3 (7B)	**-0.6522**

Table 5: Correlation between first token accuracies and question lengths (number of characters) of LLMs.

cessing abilities for complex educational content. This analysis could serve as a foundation for further research into the adaptation of language models to educational contexts, focusing on the scalability of model effectiveness across varying levels of academic complexity.

Accuracies across Number of Options. We report the accuracies of LLMs over different number of options in MalayMMLU, as depicted in Figure 4. We observe that as the number of option increases, the accuracies of the LLMs decreases, which suggest that questions with more options are more difficult to LLMs. We hypothesize this is due to as number of options increases, selecting the correct options requires a better and more thorough cognitive capability, hence poses more challenges to LLMs.

Accuracies across Question Lengths. We report the Pearson correlation coefficient between LLMs' accuracy and question length in Table 5. We observe negative correlations across all models between their accuracies and the length of questions, suggesting that as the questions are longer, LLMs are experiencing difficulties in answering the questions correctly. We conjecture that stronger models have lower correlations due to their consistent performances across different question lengths.

5.3 Analysis

Confidence on Difficult Questions. We conduct a quantitative analysis to assess the challenges posed by the MalayMMLU questions to LLMs. We define question difficulty using three criteria: (i) question length, (ii) education levels, and (iii) number of options. To explore these dimensions, we calculate correlations between LLMs' confidence scores and their correct, incorrect, and overall predictions across the dataset.

Our findings, as documented in Table 6, reveal a negative correlation between LLMs' confidence score between (i) *question length*, (ii) *education levels* and (iii) *number of options*. A negative correlation between *question length* and LLMs' confidence scores indicates that longer questions typically result in lower confidence in predictions. This trend suggests that increased textual complexity and information load may challenge the models' processing capabilities.

Further analysis in Table 6 indicates similar trends for *education levels* and *number of options*. With the increase in educational level and number of options, LLMs exhibit lower confidence scores. These results highlight that higher educational content complexity and increased decision-making demands (as indicated by more options) exacerbate the difficulty for LLMs.

These observations collectively suggest that factors such as question length, education level, and choice complexity are critical in determining the challenge level of questions for LLMs, thereby impacting their prediction confidence. Such insights underscore the importance of considering these variables in the design and training of models for educational content.

Few-Shot performance. In Figure 5, we illustrate the few-shot learning results for various LLMs using the MalayMMLU dataset. For each instance, we select examples that are specific to the subject matter of the question being addressed. For instance, only biology-related prompts are used for biology questions. Notably, the addition of few-shot examples does not appear to enhance the models' predictive capabilities. This finding aligns with those reported in CMMLU (Li et al., 2023), where few-shot prompts were found to be minimally beneficial for instruction-tuned LLMs.

This observation suggests a potential limitation in the adaptability of current instruction-tuned LLMs when faced with context-specific tasks in a

Model	Question Length			Education Level			No. of Options		
	Correct	Wrong	All	Correct	Wrong	All	Correct	Wrong	All
SeaLLM-v2.5 (7B)	0.0462	-0.0364	-0.0010	**-0.1051**	**-0.0521**	-0.1069	0.1024	-0.0149	0.0250
LLaMA-3 (8B)	-0.0460	-0.0933	-0.0905	-0.0773	-0.0498	-0.0872	-0.0887	-0.2193	-0.1771
Sailor (7B)	**-0.2038**	**-0.2560**	**-0.2450**	-0.1030	-0.0473	**-0.1168**	-0.1817	**-0.3767**	**-0.2779**
Mistral-v0.3 (7B)	-0.1302	-0.1702	-0.1701	-0.0369	-0.0426	-0.0528	**-0.1846**	-0.2666	-0.2564

Table 6: Correlation between LLMs' confidence and (i) *question length*, (ii) *education level* and (iii) *number of options*. Generally, we observe negative correlations between LLMs' confidence and all three factors.

Split / Model	Detected Malay	Detected Indonesian	Others
	45.98%	53.53%	0.48%
GPT-4	79.38	80.74	80.34
GPT-3.5	67.07	68.40	65.81
LLaMA-3 (8B)	63.33	63.66	54.70
Sailor (7B)	66.00	69.00	61.54
SeaLLM-v2.5 (7B)	65.33	66.46	55.56
Mistral-v0.3 (7B)	57.63	57.82	53.85

Table 7: Malay vs Indonesian Language: First token accuracies of various LLMs on MalayMMLU, splitted by detected language using fastText classifier.

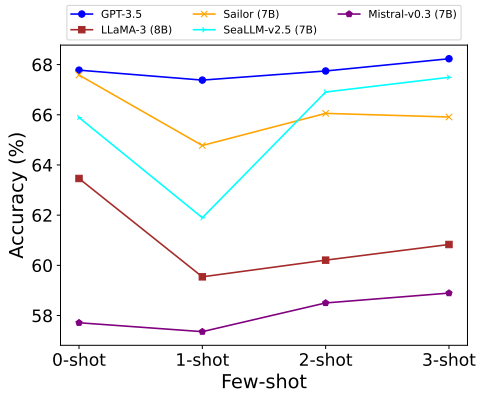

Figure 5: Few-shot results of LLMs. We observe similar performances to (Li et al., 2023).

few-shot setting. Such results highlight the need for further refinement in the training processes or model architectures to better leverage few-shot learning for specialized content.

Language Similarity. In Table 7, we present the results of applying the fastText classifier (Joulin et al., 2017) to the MalayMMLU dataset. Notably, approximately 50% of the questions in MalayMMLU are classified as Indonesian. Kargaran et al. (2023) have indicated that current language identification classifiers may suffer from contamination between data from higher-resource and lower-resource languages and face challenges in distinguishing closely related languages. Our findings affirm this perspective, underscoring the urgent need for enhanced research in language identification for closely related languages, such as Malay and Indonesian.

Further, we categorized the MalayMMLU data based on the fastText classifier's detections into Malay, Indonesian, and Other categories, and assessed their accuracies. The performance of various LLMs was found to be consistent across the fastText-detected Malay and Indonesian categories, suggesting that the models' effectiveness in handling Indonesian is likely transferable to Malay.

6 Discussion

As LLMs are gradually evolving, it is important to evaluate their performances through systematic benchmarks such as MMLU, which sheds light in understanding LLMs cognitive ability. Although being superior in various benchmarks, LLMs often struggle to comprehend the local cultures and low-resource languages, due to the scarcity of such data in their pretraining dataset. As reported in Table 4, GPT-4 is the only LLM that scores 80%, highlighting the need for improving LLMs in the low-resource languages regime, specifically for Malay.

We highlight the similarity between Indonesian and Malay (with lexical similarity of ∼90%), and suggest the performance of LLMs are likely transferable across similar language families. We further conjecture such a finding could be potentially helpful for training LLMs with low-resource languages, by pretraining on a similar, resourceful language.

7 Conclusion

This paper introduces MalayMMLU, the first multitask dataset specifically designed the Malay language, a low-resource language. MalayMMLU offers a systematic evaluation of LLMs in relation to the Malaysian educational curriculum. These results underscore the necessity for further research and development in Malay language processing. It is our hope that MalayMMLU will poise to have a substantial impact on the growth and enrichment of the Malay language, fostering advancements in natural language understanding and technology tailored to the needs of Malay-speaking communities.

Limitation

We discuss several limitations of our MalayMMLU benchmark as follows: (i) absence of multimodal questions, (ii) lack of essay-format questions, and (iii) exclusion of local colloquial variations such as the Kelantan-Malay dialect.

Firstly, we excluded all questions that required multimodal content such as images, videos, or audio to focus solely on text-based evaluations. This decision limits our ability to assess how well LLMs handle multimedia information, which is increasingly relevant in real-world applications. Secondly, MalayMMLU does not include essay-format questions, which are critical for evaluating LLMs' capabilities in generating extended text and engaging in deeper, more comprehensive language tasks. Lastly, the benchmark does not incorporate local colloquialisms, resulting in a less nuanced understanding of LLM performance when dealing with dialect-specific or culturally nuanced content. This exclusion could impact the effectiveness of LLMs in fully grasping the linguistic diversity within the Malaysian context.

Ethical Consideration

MalayMMLU is designed strictly for research purposes to advance the study of Malay, a low-resource language. It is important to note that our experimental results specifically represent the performance of LLMs on our dataset. We also want to highlight that our dataset may not accurately reflect the performance of LLMs on real-world examination questions, which often include multimodal elements and essay formats. This limitation should be considered when generalizing the findings to broader applications.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

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

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *Preprint*, arXiv:2404.03608.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Nankai Lin, Sihui Fu, Shengyi Jiang, Gangqin Zhu, and Yanni Hou. 2018. Exploring lexical differences between indonesian and malay. In *2018 International Conference on Asian Language Processing (IALP)*, pages 178–183.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Asmah Haji Omar. 2001. The malay language in malaysia and indonesia: From lingua franca to national language. *The Aseanists ASIA, II*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Bali Ranaivo-Malancon. 2006. Automatic identification of close languages – case study: Malay and indonesian. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 2.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. 2024. Mallam – malaysia large language model. *Preprint*, arXiv:2401.14680.

A Appendix

This Appendix provides additional details and experimental results to support the main submission. We begin by providing the sample questions from MalayMMLU and IndoMMLU, to highlight the similarities between the two languages in Section A.1. We then include the descriptions and data distributions of MalayMMLU in Section A.2. In Section A.3, we report additional results on MalayMMLU, including the full answer accuracies, and result breakdowns of selected LLMs on different subjects. Next, we display the few-shot prompt template in Section A.4. Lastly, we depict the model artifacts used in our experiments, in Section A.5.

A.1 Sample Questions

In Figure 6, we display sample questions from both MalayMMLU (left) and IndoMMLU (right). We observe significant similarities between both languages.

MalayMMLU Malay language (Form 5)	IndoMMLU Indonesian language (Kelas XII SMA)
Tukarkan struktur ayat pasif dalam bahasa klasik kepada bahasa moden: Bahasa klasik: Maka oleh diparang oleh Hang Tuah kepada orang mengamuk itu berbelah dua. A. Adapun memarangnya Hang Tuah kepala dua orang yang mengamuk itu dan terbelah dua. B. Terbelah dua kepala dua orang yang mengamuk itu oleh Hang Tuah. **C. Lalu Hang Tuah memarang kepala dua orang yang mengamuk itu dan terbelah dua.** D. Hang Tuah memarang kepala dua orang yang mengamuk itu dan terbelah dua.	Penyerahan barang yang tepat waktu dan bentuk layanan lainnya menjadi dominan sangat berpengaruh pada reputasi dan bonaviditas bisnis mereka. Kata dominan, reputasi, dan bonaviditas dalam kalimat itu mengandung makna... A. dapat dipercaya, nama baik, kejujuran B. sangat menentukan, nama baik, kejujuran C. berkuasa, berbuak baik, dapat dipercaya D. sangat menguasai, perbuatan baik, jujur **E. berpengaruh, nama baik, menentukan**

Figure 6: Sample questions of Malay from MalayMMLU (left) and Indonesian from IndoMMLU (right).

A.2 Data Statistics

In this section, we provide the detailed descriptions and the number of questions according to each subject in Table 8.

A.3 Additional Results

In Table 9, we report the full answer accuracies of multiple LLMs. Additionally, we report the breakdown of first token accuracies of GPT-3.5, GPT-4, LLaMA-3, Sailor, SeaLLM and Mistral, in Table 10, 11, 12, 13, 14 and 15 respectively.

A.4 Few-Shot Prompt

In this section, we display the few-shot prompt template used in our experiments, as shown in Table 16.

A.5 Model Artifacts

We include the open-source model artifacts from Hugging Face Hub in Table 17.

Category	Subject	Description	Number of questions
Social Science	History	Explores past events, particularly in human affairs	5515
	Geography	Studies Earth's lands, features, inhabitants, and phenomena	1163
	Local Studies	Focuses on the history, geography, and social aspects of local areas	240
Language	Malay Language	National language of Malaysia	6288
Humanities	Islam Studies	Understanding of the Islamic faith, its practices, and its impact on the world	4169
	Quran and Sunnah	Focuses on the study of the Quran and Sunnah, the primary sources of Islamic teachings	130
	Sports Science Knowledge	Studies the body's response to exercise and how sports enhance health	96
Others	Life Skills	Teaches practical skills everyday life	2920
	Principles of Accounting	Teaches financial accounting principles and reporting rules	752
	Business	Basics of buying, selling, producing, and distributing goods or services	199
	Economics	Creation, distribution, and use of goods and services, and the workings of economies	199
STEM	Chemistry	Studies the composition, structure, properties, and reactions of matter	482
	Computer Literacy	Teaches the confident and efficient use of computer applications	394
	Mathematics	Studies numbers, shapes, and patterns, and their properties and relationships	313
	Biology	Studies life and living organisms, including their structure, function, and evolution	282
	Computer Science	Studies computers and computing technologies, including programming and software development	277
	Design and Technology	Applies knowledge and skills to create innovative solutions to real-world problems	257
	Core Science	Provides a broad study of the material, living, and technological world	125
	Additional Mathematics	Provides a basis for more advanced studies in mathematics	110
	Information and Communication Technology	Covers technologies that provide access to information through telecommunications	105

Table 8: Summary of the subjects of MalayMMLU.

Model	Language	Humanities	STEM	Social Science	Others	Average
	Full Acc.	Full Acc.	Full Acc.	Full Acc.	Full Acc.	Full Acc.
GPT-4	**79.52**	**81.14**	**76.26**	**72.93**	**74.48**	**76.73**
GPT-3.5	67.33	69.65	65.04	63.28	61.98	65.44
LLaMA-3 (8B)	54.10	56.00	52.11	51.99	52.22	53.32
LLaMA-2 (13B)	44.99	46.39	40.11	41.01	39.67	42.70
LLaMA-2 (7B)	44.93	49.97	45.11	46.24	45.86	46.40
Mistral-v0.3 (7B)	56.23	58.23	55.26	55.52	55.12	56.10
Mistral-v0.2 (7B)	56.65	59.29	56.20	55.93	55.27	56.64
Sailor† (7B)	67.80	61.30	55.59	56.74	56.92	60.35
SeaLLM-v2.5† (7B)	63.23	61.87	58.25	58.27	57.45	60.07
Gemma (7B)	43.15	49.97	45.93	46.30	47.40	46.30
Gemma (2B)	44.64	50.78	48.92	47.79	49.08	47.85
Qwen-1.5 (7B)	55.39	55.79	51.99	50.68	52.27	53.24
Qwen-1.5 (4B)	45.77	50.97	47.81	47.37	48.57	47.86
Qwen-1.5 (1.8B)	42.81	49.19	44.99	45.20	47.95	45.76
Komodo† (7B)	42.03	49.85	44.17	45.24	46.27	45.31
MallaM-v2† (5B)	42.06	40.16	36.10	36.34	37.08	38.62
Phi-3 (14B)	59.53	56.50	57.31	55.35	52.39	56.33
Phi-3 (3.8B)	52.47	55.63	53.50	53.17	52.17	53.29

Table 9: Zero-shot results of various LLMs on MalayMMLU. The full answer accuracies are reported. Highest scores are **bolded** and second highest scores are underlined. † denotes the LLMs that are finetuned with SEA datasets. We observe that GPT-4 achieved highest accuracies across all topics.

Subject	Primary	Secondary
Information and Communication Technology	82.86	-
Core Science	77.78	72.41
Islam	77.16	67.65
History	74.94	63.50
Design and Technology	74.73	65.66
Mathematics	73.68	55.44
Local Studies	72.50	-
Malay Language	71.54	64.03
Life Skills	69.72	65.04
Additional Mathematics	-	43.64
Agriculture	-	68.69
Automotive Technology	-	65.31
Biology	-	74.82
Business	-	73.37
Chemistry	-	59.96
Computer Literacy	-	77.66
Computer Science	-	68.95
Economics	-	65.83
Geography	-	72.40
Principles of Accounting	-	52.26
Quran and Sunnah	-	61.54
Sports Science Knowledge	-	59.38

Table 10: GPT-3.5 performance (% accuracy) across Primary and Secondary education levels by subject. "-" denotes that the subject is not available in the curriculum of the education level.

Subject	Primary	Secondary
Information and Communication Technology	92.38	-
Islam	88.15	81.90
Design and Technology	85.71	69.88
Malay Language	85.65	74.88
Life Skills	84.27	76.50
History	83.53	74.92
Local Studies	83.33	-
Core Science	77.78	82.76
Mathematics	63.16	65.31
Additional Mathematics	-	51.82
Agriculture	-	78.79
Automotive Technology	-	80.61
Biology	-	87.94
Business	-	85.43
Chemistry	-	81.33
Computer Literacy	-	86.80
Computer Science	-	75.45
Economics	-	83.92
Geography	-	81.08
Principles of Accounting	-	72.07
Quran and Sunnah	-	73.08
Sports Science Knowledge	-	73.96

Table 11: GPT-4's accuracy across primary and secondary education levels by subject. "-" denotes that the subject is not available in the curriculum of the education level.

Subject	Primary	Secondary
Information and Communication Technology	79.05	-
Islam	71.93	63.15
Local Studies	71.25	-
Design and Technology	69.23	63.86
History	68.62	60.38
Life Skills	67.14	62.67
Core Science	66.67	70.69
Malay Language	65.37	59.73
Mathematics	57.89	55.10
Additional Mathematics	-	46.36
Agriculture	-	63.64
Automotive Technology	-	62.24
Biology	-	68.44
Business	-	69.35
Chemistry	-	51.66
Computer Literacy	-	71.57
Computer Science	-	62.09
Economics	-	67.34
Geography	-	67.58
Principles of Accounting	-	49.87
Quran and Sunnah	-	55.38
Sports Science Knowledge	-	56.25

Table 12: LLaMA-3 (8B) performance (% accuracy) across Primary and Secondary education levels by subject. "-" denotes that the subject is not available in the curriculum of the education level.

Subject	Primary	Secondary
Information and Communication Technology	81.90	-
Core Science	77.78	66.38
Malay Language	76.99	67.39
Islam	73.74	65.40
History	73.15	61.68
Local Studies	72.50	-
Design and Technology	71.43	65.66
Life Skills	70.66	65.24
Mathematics	52.63	53.40
Additional Mathematics	-	46.36
Agriculture	-	72.73
Automotive Technology	-	63.27
Biology	-	68.09
Business	-	71.36
Chemistry	-	51.45
Computer Literacy	-	74.87
Computer Science	-	63.18
Economics	-	65.33
Geography	-	69.05
Principles of Accounting	-	50.53
Quran and Sunnah	-	63.85
Sports Science Knowledge	-	65.62

Table 13: Sailor (7B) performance (% accuracy) across Primary and Secondary education levels by subject. "-" denotes that the subject is not available in the curriculum of the education level.

Subject	Primary	Secondary
Information and Communication Technology	83.81	-
Islam	73.56	64.99
Malay Language	71.63	64.28
Life Skills	70.42	63.63
History	69.09	59.95
Local Studies	67.50	-
Design and Technology	60.44	64.46
Mathematics	47.37	48.30
Core Science	44.44	69.83
Additional Mathematics	-	47.27
Agriculture	-	73.74
Automotive Technology	-	70.41
Biology	-	70.57
Business	-	74.37
Chemistry	-	61.20
Computer Literacy	-	78.17
Computer Science	-	67.15
Economics	-	66.33
Geography	-	67.93
Principles of Accounting	-	54.79
Quran and Sunnah	-	58.46
Sports Science Knowledge	-	55.21

Table 14: SeaLLM-v2.5 (7B) performance (% accuracy) across Primary and Secondary education levels by subject. "-" denotes that the subject is not available in the curriculum of the education level.

Subject	Primary	Secondary
Information and Communication Technology	72.38	-
Core Science	66.67	68.10
Islam	66.30	54.78
Design and Technology	65.93	60.24
Local Studies	65.00	-
History	62.89	56.08
Life Skills	62.68	57.78
Malay Language	57.66	54.93
Mathematics	36.84	50.00
Additional Mathematics	-	39.09
Agriculture	-	67.68
Automotive Technology	-	58.16
Biology	-	60.28
Business	-	66.33
Chemistry	-	48.76
Computer Literacy	-	65.48
Computer Science	-	57.04
Economics	-	55.28
Geography	-	62.42
Principles of Accounting	-	45.35
Quran and Sunnah	-	54.62
Sports Science Knowledge	-	55.21

Table 15: Mistral-v0.3 (7B) performance (% accuracy) across Primary and Secondary education levels by subject. "-" denotes that the subject is not available in the curriculum of the education level.

0-shot	Multi-shot
Berikut adalah soalan aneka pilihan tentang [Subject]. Sila berikan jawapan sahaja. [Question] Jawapan:	Berikut adalah soalan tentang [Subject]. [Example question 1] Jawapan: [Answer 1] [Example question 2] Jawapan: [Answer 2] [Example question 3] Jawapan: [Answer 3] [Question] Jawapan:

Table 16: The prompt template for MalayMMLU in zero-shot and multi-shot setting. On the right, we show an example of prompt template in 3-shot setting.

Models (#parameters)	Source
GPT-4	gpt-4-turbo-2024-04-09
GPT-3.5	gpt-3.5-turbo-0125
LLaMA-3 (8B)	meta-llama/Meta-Llama-3-8B-Instruct
LLaMA-2 (13B)	meta-llama/Llama-2-13b-chat-hf
LLaMA-2 (7B)	meta-llama/Llama-2-7b-chat-hf
Mistral-v0.3 (7B)	mistralai/Mistral-7B-Instruct-v0.3
Mistral-v0.2 (7B)	mistralai/Mistral-7B-Instruct-v0.2
Sailor (7B)	sail/Sailor-7B-Chat
SeaLLM-v2.5 (7B)	SeaLLM-7B-v2.5
Phi-3 (14B)	microsoft/Phi-3-medium-4k-instruct
Phi-3 (3.8B)	microsoft/Phi-3-mini-4k-instruct
Qwen-1.5 (7B)	Qwen/Qwen1.5-7B-Chat
Qwen-1.5 (4B)	Qwen/Qwen1.5-4B-Chat
Qwen-1.5 (1.8B)	Qwen/Qwen1.5-1.8B-Chat
Gemma (7B)	google/gemma-7b-it
Gemma (2B)	google/gemma-2b-it
Komodo (7B)	Yellow-AI-NLP/komodo-7b-base
MallaM-v2 (5B)	mesolitica/mallam-5b-20k-instructions-v2

Table 17: All the models used in this study were sourced from Hugging Face Hub except GPT-3.5 and GPT-4.