SinhalaMMLU: A Comprehensive Benchmark for Evaluating Multitask Language Understanding in Sinhala

Ashmari Pramodya[1], Nirasha Nelki[2], Heshan Shalinda[3], Chamila Liyanage[2], Yusuke Sakai[1], Randil Pushpananda[2], Ruvan Weerasinghe[3], Hidetaka Kamigaito[1], Taro Watanabe[1]

[1]Nara Institute of Science and Technology (NAIST), Japan

[2]University of Colombo School of Computing (UCSC), Sri Lanka

[3]Informatics Institute of Technology (IIT), Sri Lanka

{pussewala.ashmari.ow4, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

{2021cs127@stu, rpn@, cml@}ucsc.cmb.ac.lk, {heshan.20230591, ruvan.w}@iit.ac.lk

Abstract

Large Language Models (LLMs) demonstrate impressive general knowledge and reasoning abilities, yet their evaluation has predominantly focused on global or anglocentric subjects, often neglecting low-resource languages and culturally specific content. While recent multilingual benchmarks attempt to bridge this gap, many rely on automatic translation, which can introduce errors and misrepresent the original cultural context. To address this, we introduce SinhalaMMLU, the first multiple-choice question answering benchmark designed specifically for Sinhala, a low-resource language. The dataset includes over 7,000 questions spanning secondary to collegiate education levels and is aligned with the Sri Lankan national curriculum. It covers six domains and 30 subjects, encompassing both general academic topics and culturally grounded knowledge. We evaluate 26 LLMs on SinhalaMMLU and observe that, while Claude 3.5 sonnet and GPT-4o achieve the highest average accuracies at 67% and 62% respectively, overall model performance remains limited. In particular, models struggle in culturally rich domains such as the Humanities, revealing substantial room for improvement in adapting LLMs to low-resource and culturally specific contexts.

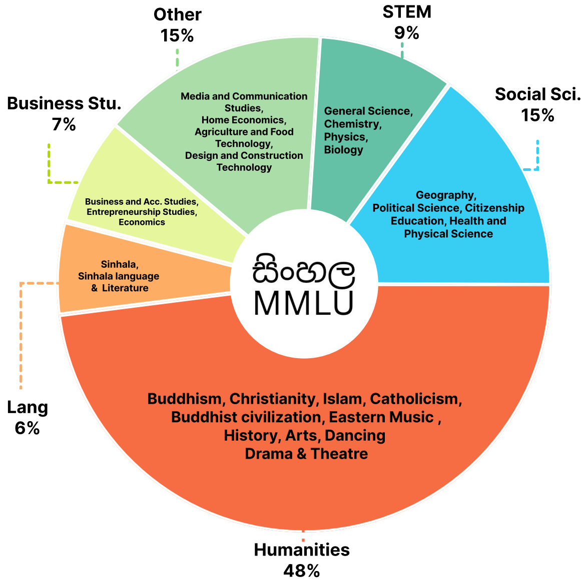

Figure 1: Distribution of the number of questions across different domains in SinhalaMMLU.

1 Introduction

The introduction of large language models (LLMs) has led to unprecedented improvements in natural language processing (NLP) and artificial intelligence, reshaping the field with their revolutionary capabilities (OpenAI et al., 2024; Anthropic, 2024; Touvron et al., 2023). Despite multilingual capabilities in LLMs, significant performance gaps persist between English and other languages (Liu et al., 2024). This performance gap is particularly pronounced for languages using non-Latin scripts such as Sinhala. One critical factor is the lack of high quality benchmarks for non-English languages, particularly those that are low resourced. The majority of current evaluation frameworks remain English-centric (Hendrycks et al., 2021), creating significant assessment gaps for linguistically diverse and low-resource languages. Many benchmarks claiming multilingual support are simply translated versions (Bandarkar et al., 2024; Singh et al., 2025) of content originally designed for English speakers. These often include content that assumes knowledge of the American legal system or requires familiarity with English-specific cultural references and colloquialisms. Despite appearing multilingual through translation, these benchmarks fail to reflect the cultural context (Ji et al., 2023) valued by native speakers.

To address the lack of existing benchmarks for the Sinhala NLP community and to better assess how LLMs perform on Sinhala, we introduce Sin-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32943–32961
November 4-9, 2025 ©2025 Association for Computational Linguistics

halaMMLU, a comprehensive evaluation framework comprising over 7,000 questions across 6 domains and 30 subjects reflecting both general and culturally specific knowledge. SinhalaMMLU is designed by collecting questions from various national and provincial exams, without relying on translations. These questions include culturally relevant subjects such as Sri Lankan history, arts, drama and theatre, indigenous dancing, oriental music, and Sinhala language, alongside traditional academic subjects like science.

We evaluate 26 different LLMs, a mix of closed proprietary and open-source. Our results show significant room for improvement, with Claude-sonnet 3.5 scoring the highest at 67.65%. We also explore how performance scales with the number of parameters, finding significant improvements as model size increases. Our domain-wise analysis reveals that models underperform in culturally relevant areas, such as Humanities and Language, as well as in general subjects within the STEM domain. All data set details are publicly available on GitHub[1] and Hugging Face [2].

2 Related work

2.1 Benchmarks for LLMs

In an era where different LLMs are emerging day by day, benchmarks are crucial to understanding their power and performance. Benchmarks like GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), and SQuAD (Rajpurkar et al., 2016) are being used as evaluation frameworks for natural language understanding (NLU) tasks, including reading comprehension and question-answering. Eventually, multilingual evaluation frameworks emerged like XGLUE (Liang et al., 2020) and XTREAM-R (Ruder et al., 2021). However, these benchmarks only assess the linguistic skills of the language rather than language understanding.

MMLU (Hendrycks et al., 2021) is one of the prominently used benchmarks for LLMs. It evaluates the LLMs via multiple-choice questions belonging to different subjects and domains to evaluate the knowledge acquisition of LLMs. Since MMLU consists of English questions, there are some attempts to create similar benchmarks for other languages, including Arabic (Koto et al., 2024), Chinese (Li et al., 2024), Turkish (Yüksel et al., 2024), Indonesian (Koto et al., 2023), Korean (Son et al., 2025), and Persian (Ghahroodi et al., 2024). Some MMLU benchmarks have been constructed for low-resourced languages, notably African (Adelani et al., 2025), Malay (Poh et al., 2024), and Basque (Etxaniz et al., 2024). The MILU benchmark that consists of 11 Indic languages, stands out among the language-specific MMLU (Verma et al., 2025). In an effort to establish a multilingual benchmark, the Include benchmark (Romanou et al., 2024) has achieved a remarkable feat as a benchmark by creating an impressive QA dataset that encompasses a variety of exams from various languages.

2.2 Sinhala Language

Sinhala is an Indo-Aryan language spoken by over 17 million people in Sri Lanka, where it serves as one of the two official languages. Despite its sizable community, Sinhala remains relatively under-resourced in terms of language technology and NLP tools, though there have been some attempts for it.

The existing textual datasets for Sinhala have been curated for different purposes like text summarization (Hasan et al., 2021; Hewapathirana et al., 2024), and text classification (Hettiarachchi et al., 2024; Warusawithana et al., 2022; Ranasinghe et al., 2024). The Aya Dataset (Singh et al., 2024), a human-annotated multilingual instruction fine-tuning dataset, also includes Sinhala data. However, these works have only focused on a single task or a single domain; therefore, not varied enough to consider for a multi-task benchmark for LLMs. Since the MMLU has been proposed for English, there have been attempts to create multilingual benchmarks by translating the initial dataset into other languages, including Sinhala (Singh et al., 2025). However, there is a question regarding the accuracy of the translation, as it can be challenging to translate language-specific terms. A language-specific benchmark or a multilingual benchmark should consider culture-related aspects as well, in order to ensure LLMs understand the cultural contexts that are lacking in the GlobalMMLU (Singh et al., 2025).

3 Sinhala MMLU

Motivated by the scarcity of high-quality resources for Sinhala, we introduce SinhalaMMLU, a bench-

[1]https://github.com/naist-nlp/SinhalaMMLU

[2]https://huggingface.co/datasets/naist-nlp/SinhalaMMLU

mark designed to evaluate language models in the Sinhala language. The benchmark includes multiple-choice questions contextually designed for Sinhala language, ranging over a wide range of subjects and educational levels. Following the structure and methodology of the original English MMLU benchmark, we carefully curated the dataset in alignment with the local national curriculum to ensure both relevance and cultural appropriateness.

The national educational curriculum of Sri Lanka is divided into three broad levels: primary, secondary, and collegiate education. Primary education spans five years (Grades 1-5) for children aged 5 to 9, focusing on foundational subjects. However, we excluded this level from Sinhala-MMLU, as the questions are generally simple (e.g., drawing, filling blanks, or writing), and are not suited for multiple-choice evaluation. Secondary education is divided into two phases: junior secondary (Grades 6 to 9, ages 10 to 13) and senior secondary (Grades 10 to 11, ages 14 to 16), culminating in the General Certificate of Education Ordinary Level (GCE O-Level) examination, where students take six main subjects and three additional subjects. The collegiate level comprises Grades 12-13 (ages 17-19) and concludes with the GCE Advanced Level (A-Level) examination. There are five academic streams at this level: Arts, Commerce, Biological Science, Physical Science, and Technology. Students select at least three subjects within their chosen stream, and the national examination includes these core subjects along with a General English test and a Common General Test. The A-Level exam serves as the primary qualification for university entrance in Sri Lanka.

All questions were prepared in alignment with the official curriculum set by the Ministry of Education (MOE), Sri Lanka. By aligning Sinhala-MMLU with educational standards, we aim to provide a rigorous and contextually grounded benchmark to evaluate the performance of LLMs in Sinhala across educational levels. This allows us to systematically assess model understanding of local knowledge, identify linguistic or subject-specific weaknesses, and encourage the development of culturally relevant and educationally aligned language models for low-resource languages.

Group	# Questions	# Chars	
		Question	Answer
Easy	1893	59.08	16.77
Medium	2585	100.66	24.79
Hard	2566	116.40	27.53
STEM	629	157.82	27.42
Social Science	1084	141.80	22.34
Humanities	3419	93.91	22.24
Language	397	74.19	25.65
Business Studies	477	173.39	32.99
Other	1038	108.58	28.24

Table 1: Total number of questions, Average question and answer length (in characters) for each difficulty level and domain.The overall question count is 7044.

3.1 Data Preparation

Our primary data source was e-thaksalawa[3], the public official e-learning platform developed by the Sri Lankan MOE. This platform provides free access to a wide range of educational materials, including past exam papers, lecture notes, interactive textbooks, video lessons, and multiple-choice questions (MCQs) with answer keys curated by domain experts. We collected MCQs aligned with the national curriculum from government-focused papers, specifically from the GCE O-Level, GCE A-Level, and provincial examination papers for Grades 6–8. These exams include multiple-choice sections covering various academic subjects. In addition to e-Thaksalawa, we used other educational websites that host government-released past papers and marking schemes.

To curate the data, we employed four annotators with undergraduate or higher educational qualifications. Over a two-month period, the annotators manually extracted MCQs from PDF documents (after OCR), while quiz-based content was scraped directly from the web platform[4]. Annotators were instructed to include only questions that were in multiple-choice format with clearly defined answer options and an identified correct answer. Questions containing multimodal content, such as images, audio, or video, were excluded. Furthermore, we excluded Mathematics questions because they primarily involve symbolic reasoning and are not usually provided as MCQ questions. More than

[3] https://www.ethaksalawa.moe.gov.lk/En/index.php

[4] https://www.e-thaksalawa.moe.gov.lk/moodle/course/view.php?id=636

Domain	Subject
Humanities	History Drama and Theatre Dancing Eastern Music Arts Buddhism Catholicism Christianity Islam Buddhist Civilization Oriental Music History of Sri Lanka Dancing Indigenous
Social Science	Citizenship Education Health and Physical Science Geography Political Science
STEM	Physics Chemistry Biology Science
Language	Sinhala Language and literature
Business Studies	Business and Accounting Studies Entrepreneurship Studies Economics
Other	Home Economics Biosystems Technology Communication and Media Studies Design and Construction Technology Agriculture and Food Technology

Table 2: Subjects categorized by domain in the Sinhala-MMLU dataset.

65% of the collected data was sourced from PDFs.

3.2 Data Formatting

Each MCQ in SinhalaMMLU follows a consistent structure composed of a question stem, a set of four or five choices, and a single correct answer. The dataset includes both fill-in-the-blank style questions and direct queries. Along with each question, we recorded metadata such as the subject, difficulty level mapped to school grade, original source and year, and the province, especially in the case of government-issued or provincial papers. For chemical formulae and mathematical expressions, we use a 50:50 mixture of LATEX and plain text, where plain text was only allowed if an expression is commonly used and not prone to ambiguity. To ensure content uniqueness and eliminate redundancy, we removed exact duplicate questions with identical answer sets and applied cosine similarity analysis. Questions with a similarity score exceeding 95% were flagged and removed from the dataset. Example questions are illustrated in Figure 2.

3.3 Difficulty Classification

We classified question difficulty as follows. Questions from Grades 6–8 were labeled as **easy**. Most of these had four answer choices, although a subset of geography questions (44 in total) included only three. To maintain format consistency, we added a fourth option using Claude 3.7 Sonnet, which was not used during model evaluation with our dataset. Questions from Grades 9–11 were considered **medium** in difficulty, all of which contained four answer options. Questions from Grades 12–13 were labeled as **hard** and typically featured five answer choices. This classification mirrors the increasing academic depth and complexity expected at higher grade levels and provides a clear basis for evaluating model performance across educational stages.

3.4 Data Distribution

Table 1 summarizes the dataset statistics, including the average question length across 7,044 questions. Each subject has at least 104 questions, which we split into few-shot set with 3 questions, and test set with more than 100 questions. As shown in Table 1 and accompanying Figure 1, our SinhalaMMLU includes the following domains: “Humanities” which is deeply rooted in Sri Lankan cultural and historical context; “STEM”,“Social Science”,“Business Studies”, “Language”, and “Other”, For a detailed breakdown of the subjects within each domain and their difficulty levels in Table 2, refer to Appendix A.

4 Experimental Setup

Models We evaluated 26 recent state-of-the-art multilingual LLMs of different sizes in zero-shot and few-shot settings. These include both small and large open models as well as closed models. For open source models, we include Cohere4AI’s Aya Expanse[5], LLaMA-3 (Touvron et al., 2023), Qwen (Qwen et al., 2025), and Mistral (Jiang et al., 2024), along with their chat versions. To optimize memory efficiency during inference, we applied 4-bit NF4 quantization with double quantization and bfloat16 computation. For closed models, we used GPT-4o (OpenAI et al., 2024) and Claude

[5]https://hf.co/blog/aya-expanse

STEM : Physics A/L	Social Science : Citizenship Edu. O/L	Humanities : History O/L	Language : Sinhala Grade 6
ප්‍රකාශ ඉලෙක්ට්‍රෝන නිපදවීමට හැකියාව ඇති ලෝහ පෘෂ්ඨයක් මත ඒකවර්ණ ආලෝක කදම්බයක් පතිත වේ. ආලෝකයේ සංඛ්‍යාතය මෙම ලෝහය සඳහා කපා හරින සංඛ්‍යාතයට වඩා වැඩි නම්, ලෝහ පෘෂ්ඨයෙන් විමෝචනය වන ප්‍රකාශ ඉලෙක්ට්‍රෝන සංඛ්‍යාව සමානුපාතික වනුයේ, 1. ප්‍රකාශ ඉලෙක්ට්‍රෝනයක චාලක ශක්තියෙහි පරස්පරයට ය. 2. ලෝහයේ කාර්ය ශ්‍රිතයට ය. 3. පතිත. ආලෝකයේ සංඛ්‍යාතයට ය. **4. ලෝහ පෘෂ්ඨය මත වදින ෆෝටෝන සංඛ්‍යාවට සය.** 5. එක් ෆෝටෝනයක ශක්තියට ය.	ශ්‍රී ලංකාවට 'රාජ්‍ය මන්ත්‍රණ සභා' පාලන ක්‍රමය හඳුන්වා දුන්නේ, 1. මැකලම් ආණ්ඩුක්‍රම ප්‍රතිසංස්කරණ මගිනි. **2. ඩොනමෝර් ආණ්ඩුක්‍රම ප්‍රතිසංස්කරණ මගිනි.** 3. සෝල්බරි ආණ්ඩුක්‍රම ප්‍රතිසංස්කරණ මගිනි. 4. මැනිං ආණ්ඩුක්‍රම ප්‍රතිසංස්කරණ මගිනි.	පහත දැක්වෙන ඓතිහාසික සිදුවීම් ඇසුරෙන් පහත ප්‍රශ්නයට පිළිතුරු සපයන්න. A - රුවන්වැලි මහා සෑය ඉදි කිරීම B - ලම්භකර්ණ රාජ වංශයේ ආරම්භක පාලකයා වීම C - ත්‍රිපිටකය ග්‍රන්ථාරූඪ කිරීම D - ප්‍රථම සංවිධානාත්මක සටන් ව්‍යාපාරය ඇරඹීම E - පළමුවරට විශාල ප්‍රමාණයේ වැව් ඉදි කිරීම වසභ රාජ්‍ය සමයට අයත්වන සිදුවීම් දැක්වෙන පිළිතුර කුමක් ද? 1. A හා C 2. A හා D 3. B හා C **4. B හා E**	ප්‍රබුද්ධ කාව්‍ය ග්‍රන්ථයේ රචකයා වන්නේ 1. ගුණදාස අමරසේකර ය. 2. සිරි ගුනසිංහ ය. 3. පී. බී. අල්විස් පෙරේරා ය. **4. මහගමසේකර ය.**
A beam of monochromatic light is incident on a metal surface which is capable of producing photoelectrons. If the frequency of the light is above the cut-off frequency of the metal, the number of photoelectrons ejected from the metal surface is proportional to the 1. reciprocal of the kinetic energy of a photoelectron. 2. work function of the metal. 3. frequency of the incident light. **4. number of photons that hits the metal surface.** 5. energy of a single photon.	'Rajya Mantrana Saba' system of government was introduced to Sri Lanka 1. by McCallum Constitutional Reforms. **2. by Donoughmore Constitutional Reforms.** 3. by Soulbury Constitutional Reforms. 4. by Manning Constitutional Reforms.	Answer the following question based on the historical events listed below. A - Construction of Ruvanväli mahā säya B - First ruler of the Lambakarna dynasty C - Committing the Tripitaka (canon) to writing D - Initiation of the first organised fighting movement E - Construction of large scale tanks for the first time What answer includes the incidents belonging to the reign of Vasabha? 1. A and C 2. A and D. 3. B and C. **4. B and E**	The author of the poetry book 'Prabuda' is 1. Gunadasa Amarasekara. 2. Siri Gunasinghe. 3. P. B. Alwis Perera. **4. Mahagamasekara.**

Figure 2: Examples of questions from 4 domains with answers in **bold**. English translations are provided below each question. The table also includes a sample suboption-type question from the Humanities domain.

(Anthropic, 2024). The details of these models are provided in the Appendix B.

Evaluation We evaluated LLMs by accuracy. Following previous studies (Koto et al., 2023; Li et al., 2024), for open-source models, we determine the answer based on the highest probability among all possible options. For closed-source models, we determine the answer based on the first token generated in the text using a regular expression.

Prompt For the SinhalaMMLU evaluation, we used two distinct prompt templates, one that specifies the subject domain and another that omits this information. Both follow the instruction format proposed by Hendrycks et al. (2021), with prompt instructions given in the same language as the question. More details in Appendix B.1.

5 Results and Discussions

This section presents the evaluation results on the SinhalaMMLU dataset. To assess the impact of subject-specific information in prompts, we compared two variants: one incorporating the subject name and one without. We observed that including the subject name generally improves performance across models. Therefore, for all subsequent experiments, we report results using prompts that include subject information. Appendix C provides a detailed comparison.

Results by model Table 3 summarizes the full results of all the models, grouped by domain. Closed models, notably Claude 3.5 Sonnet and GPT-4o, exhibit superior performance across all domains, achieving average accuracies of 67.65% and 62.95%, respectively. These models consistently outperform open-source counterparts, indicating a higher level of generalization and understanding. Among open-source models, Qwen2.5-72B-chat and Qwen2.5-32B-chat stand out, with average accuracies of 41.18% and 36.47%, respectively. Their performance suggests that larger model sizes and instruction tuning contribute positively to handling diverse and complex tasks. While within the LLaMA-3 model family, LLaMA-3.1-70B-chat is the only variant that achieves comparatively higher performance. In contrast, smaller models like LLaMA-3.2-1B and Mistral-7B variants show lower performance, with average accuracies hovering around 22%. This disparity underscores the importance of model scale and training data diversity in achieving higher accuracy. Figure 3 illustrates that the scaling law is still present, as it is evident that models with large sizes perform better than smaller ones. A similar trend across model families is also observed in Winata et al. (2025).

Results by Domain Models consistently underperformed in culturally grounded areas, with Humanities and Language achieving average scores of 66.15% and 62.37% respectively with Claude. These results reflect the inherent challenges these models face when navigating cultural context and idiomatic expressions that are deeply embedded in local knowledge. Similarly, performance in STEM was also limited, with an average of 61.40%, likely due to the difficulty of handling domain-specific terminology and complex conceptual rea-

Model	Humanities	Language	Social Science	STEM	Business Studies	Other	Average
Claude-3-5-sonnet	**66.15**	**62.37**	**77.55**	**61.40**	**73.12**	**65.58**	**67.65**
Claude-3-haiku	41.01	43.81	50.57	35.34	44.09	39.64	42.14
GPT-4o	62.02	51.29	71.32	60.59	67.53	61.44	62.95
Qwen2.5-7B	22.27	21.65	25.66	18.57	20.22	21.70	22.20
Qwen2.5-7B-chat	28.02	25.00	28.30	28.50	27.10	23.67	27.23
Qwen2.5-32B	28.85	28.35	35.28	26.38	25.16	27.51	29.15
Qwen2.5-32B-chat	36.70	32.73	38.96	38.60	34.84	34.02	36.47
Qwen2.5-72B	34.69	28.87	40.57	37.95	28.82	34.52	35.14
Qwen2.5-72B-chat	39.84	38.92	45.85	45.60	41.29	38.86	41.18
Llama-3.2-1B	22.18	21.65	25.66	18.40	19.78	21.70	22.12
Llama-3.2-1B-chat	22.27	21.65	25.75	18.73	19.78	21.70	22.20
Llama-3.2-3B	22.18	22.68	25.57	18.40	19.57	21.70	22.14
Llama-3.2-3B-chat	22.27	21.65	25.66	18.57	19.35	21.70	22.14
Llama-3-8B	22.48	21.65	26.32	19.54	20.86	21.50	22.51
Llama-3-8B-chat	22.87	21.65	26.98	19.38	19.78	23.18	22.96
Llama-3-70B	22.48	23.45	24.95	22.48	19.78	22.09	22.65
Llama-3-70B-chat	27.15	20.36	26.32	23.78	20.65	22.49	25.21
Llama-3.1-8B	22.90	23.45	27.17	19.54	23.44	23.37	23.39
Llama-3.1-8B-chat	25.29	25.52	29.06	22.31	21.08	24.95	25.28
Llama-3.1-70B	22.19	22.94	26.42	19.54	19.78	21.60	22.40
Llama-3.1-70B-chat	27.24	26.80	25.94	23.78	26.67	25.25	26.37
Llama-3.3-70B-chat	24.54	23.45	27.26	23.13	25.81	22.88	24.61
Mistral-7B-chat	21.55	22.94	26.92	22.36	20.00	20.81	22.28
Mistral-7B	22.18	21.65	25.66	18.40	19.78	21.65	22.12
aya-expanse-8B	22.78	24.23	25.09	19.71	20.43	21.70	22.62
aya-expanse-32B	23.97	25.00	30.00	25.08	23.66	24.65	25.14
Avg	27.37	26.10	31.79	27.16	27.78	28.07	28.38

Table 3: Zero-shot performance (% accuracy) of LLMs across the six domains.

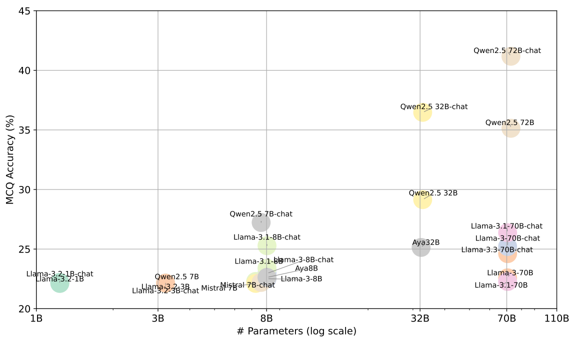

Figure 3: Individual models plotted by parameter count (x-axis, log scale) against multiple-choice question (MCQ) accuracy on the SinhalaMMLU benchmark (y-axis).

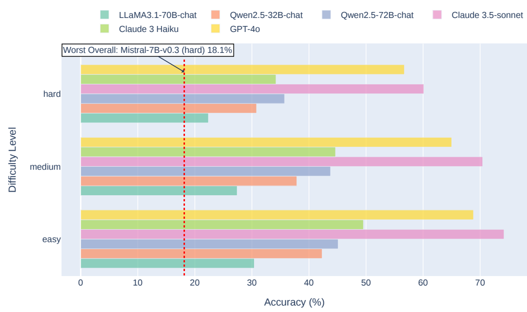

Figure 4: LLM performance across different difficulty levels.

soning. However, Social Sciences demonstrated markedly stronger performance, achieving 77.55% accuracy. This success can be attributed to the predominantly factual and descriptive nature of subjects such as Geography, Citizenship Education, and Health, which rely on standardized terminology that is well represented in the training data of most multilingual LLMs. Appendix C.2 provides more details.

Results across different difficulty levels Figure 4 presents the average scores of the top-performing models (Claude 3.5 Sonnet, GPT-4o, Claude 3 Haiku, Qwen-72B-Chat, Qwen-32B-Chat, and LLaMA-3.1-70B-Chat) across the difficulty levels defined in Section 3.3. The results in-

Model	0-shot (%)	3-shot (%)
Qwen2.5-72B-Chat	41.18	41.24
Qwen2.5-72B	35.14	35.20
Qwen2.5-32B-Chat	36.47	37.54
Qwen2.5-32B	29.15	29.68
LLaMA-3.1-70B-chat	26.37	25.05
LLaMA-3.1-70B	22.40	22.89

Table 4: Accuracy comparison of models under 0-shot and 3-shot settings

Model	Negations		Suboptions	
	w/	w/o	w/	w/o
claude-3-5-sonnet	58.36	67.66	57.45	68.72
GPT-4o	59.59	62.91	53.14	64.05
claude-3 Haiku	32.65	42.35	28.11	43.60
Qwen2.5-72B-Chat	25.30	41.82	34.10	41.91
Qwen2.5-32B-Chat	24.48	37.12	26.88	37.47
LLaMA-3.1-70B-chat	22.45	26.00	20.73	26.96

Table 5: Average accuracy classified by questions with and without negation expressions and Suboptions.

Model	5 options	4 options	Δ
claude-3-5-sonnet	60.11	66.13	+6.02
GPT-4o	56.71	62.17	+5.46
claude-3 Haiku	34.21	39.44	+5.23
Qwen2.5-72B-Chat	35.72	40.95	+6.74
Qwen2.5-32B-Chat	30.81	34.73	+3.92
LLaMA-3.1-70B-chat	22.38	26.81	+4.43

Table 6: Performance comparison on the original SinhalaMMLU and its culturally grounded subset.

dicate that Sinhala questions at the collegiate level (hard), which typically include five answer options, pose greater challenges for all models.

Few shot vs zero shot As shown in Table 4, few-shot prompting results in limited or inconsistent improvements across top-performing open-source models. While base models such as Qwen-32B show marginal gains, instruction-tuned models like LLaMA-3.1-70B-chat exhibit performance degradation, likely due to their optimization for dialogue-based tasks rather than in-context learning. These findings are consistent with prior work by Verma et al. (2025), which also observed similar limitations in instruction-tuned models under few-shot settings. More detailed in Appendix C.4.

Is negations challenging? Negation is frequently used in Sri Lankan school exam questions to increase question difficulty and evaluate students' capacity for reasoning. In the domain of NLP, prior work has similarly demonstrated that negation poses significant challenges to model performance (Truong et al., 2023). To examine its impact within our benchmark, we adopt a simple string-matching method to identify and analyze questions containing negation. We utilize specific negation phrases to identify questions containing negations in Sinhala. These include: නොගැළපේ (Not compatible), අසත්‍ය (False), වැරදි (wrong), සාවද්‍ය (False). In Table 5, most models perform less effectively on questions containing negation compared to those without, including Claude 3.5 Sonnet. GPT-4o is the only model that maintains robust performance. This pattern aligns with previous findings reported by Li et al. (2024).

Impact of Suboption Questions Similar to negation-based questions, suboption questions are another type used in exams to make questions more challenging. These questions typically consist of a main question followed by multiple subparts (e.g., labeled A, B, C, D). Common formats include selecting the correct sequence of events, matching items across suboptions, or determining whether one statement correctly explains another. These questions account for about 9.8% of the data. They often demand models to engage in deeper reasoning, including multi-step inference, comparison of related statements, and understanding of logical or causal relationships. According to Table 5, we observed that all models have weaker performance on these questions.

6 Do Hard Questions Get Easier with Fewer Options?

In Sri Lankan GCE A-Level exams, which fall under the hard difficulty level in our classification, multiple choice questions typically include five answer choices, unlike the standard four used in most evaluation benchmarks. To investigate whether reducing the number of options improves model performance, we compared accuracy between the original 5-option format and a modified 4-option version, in which one incorrect option was randomly removed. As shown in Table 6, all models improved with fewer options. Claude 3.5 Sonnet and GPT-4o saw gains of +6.02 and +5.46 points, respectively. However, even with 4 choices, performance remained lower than on medium-difficulty questions, suggesting that reduced options do not fully offset the complexity of hard-level content. To

Model	4 Options (Corr)	5 Options (Corr)
QWEN2.5-72B-CHAT	0.6835	0.7588
QWEN2.5-32B-CHAT	−0.0045	0.0696
LLAMA-3.1-70B-CHAT	−0.0602	−0.5529

Table 7: Correlation between model confidence and accuracy under 4-option and 5-option settings.

Data set	Score (%)
SinhalaMMLU (Ours)	97.30
GlobalMMLU-si (Singh et al., 2025)	71.07

Table 8: Naturalness ratings: GlobalMMLU-si vs. SinhalaMMLU STEM questions. We applied a linear transformation to convert the 5-point naturalness scores into a 100-point scale for ease of interpretation.

further explore model behavior, we analyzed confidence scores across both settings. Interestingly, while confidence scores tend to be higher in the 4-option setup (Table 7), correlation between confidence and accuracy remains weak or inconsistent across models. For instance, Qwen2.5-72B-chat shows moderate positive correlation in the five-option format. These findings suggest that reducing the number of answer choices can simplify the task for LLMs and lead to better performance, particularly on harder, culturally grounded questions. However, improved accuracy does not always align with model confidence.

7 Do Translated STEM Questions Match the Quality of Native Ones?

Translation is the simplest way to scale multilingual benchmarks. Attempts have been made to translate MMLU to Sinhala (Singh et al., 2025). A key challenge in adapting those benchmarks to low-resource languages is the accurate translation of domain-specific terminology, particularly in scientific subjects. In Sinhala, scientific concepts are conveyed through precise and contextually appropriate vocabulary that reflects standardized usage in the national curriculum. Literal or machine-translated versions from English frequently fail to capture these nuances, resulting in awkward or unnatural phrasing that deviates from actual usage.

To assess the linguistic quality of translated versus native content, we randomly sampled 100 questions from our native Sinhala MMLU benchmark and 100 questions from a Sinhala-translated subset of the global MMLU STEM dataset, excluding mathematics and engineering. Two Sinhala-speaking annotators from the Biology stream[6] of Sri Lanka's Advanced Level (A/L) education system were asked to rate each question's linguistic naturalness on a 5-point scale, focusing solely on language fluency and terminology, independent of the correctness of the answer.

Table 8 shows the comparison of the naturalness. Native Sinhala content scored significantly higher in linguistic fluency (97.3) compared to translated questions (71.07). Our analysis reveals that while translation offers a scalable approach, it often fails to capture the precise technical terms used in Sinhala STEM subjects, resulting in awkward or misleading phrasing. For instance, "plasmolysis" was incorrectly translated as ප්ලාස්මොලිසිස් (plāsmolisis) instead of the proper technical term ප්ලාස්මච්ඡේදනය. (plāsmacchēdanaya). Together, these results emphasize that direct translation is insufficient for creating high-quality multilingual benchmarks (Sakai et al., 2024).

8 Cultural Knowledge Analysis

To analyze how well LLMs understand Sri Lankan culture and the linguistic characteristics of the Sinhala language, we manually annotated questions that require knowledge of Sinhala vocabulary, literature, and culturally embedded concepts. Most culturally relevant items were concentrated in domains such as drama and theatre, oriental music, traditional dance, Sinhala language and literature, and Sri Lankan history.

In this section, we present the performance of selected models on these culturally grounded questions. The SinhalaMMLU cultural subset consists of a collection of 1,608 (22%) handpicked questions. Table 9 presents a performance analysis, focusing on the discrepancy between general performance and Sinhala cultural-specific questions, revealing a consistent performance drop across all models on this cultural subset. Closed-source models such as Claude 3.5 Sonnet and GPT-4o, experienced substantial drops of 28.22 and 23.83 percentage points when evaluated solely on culturally relevant questions. Interestingly, LLaMA-3.1-70B-Chat showed the smallest drop (-0.01), but its orig-

[6] In the Sri Lankan Advanced Level (A/L) education system, students specialize in one of five streams: Biology, Physical Science, Commerce, Arts, or Technology. The Biology stream includes biology and chemistry as compulsory subjects, along with either physics or agriculture as a third subject. The annotators had studied biology, chemistry, and physics under this stream.

Model	Original	Cul.	Δ
CLAUDE-3-5-SONNET	67.65	39.43	-28.22
GPT4O	62.95	39.12	-23.83
CLAUDE-3 HAIKU	42.14	28.11	-14.03
QWEN-72B-CHAT	41.18	30.03	-11.15
QWEN-32B-CHAT	36.17	28.36	-7.81
LLAMA-3.1-70B-CHAT	26.37	26.36	-0.01

Table 9: Performance comparison on the original SinhalaMMLU and and its culturally grounded subset(Cul).

Figure 5: Subject-wise accuracy of LLMs on a culturally significant subset of the SinhalaMMLU dataset.

inal performance was already significantly lower (26.37%), indicating limited overall capability. Our subject-wise analysis (Figure 5) reveals Claude 3.5 Sonnet performs particularly well in areas related to the Sinhala language and Buddhist concepts, while exhibiting weaknesses in domains such as traditional music, drama, and history. In contrast, GPT-4o demonstrates complementary strengths, with comparatively better performance in subjects related to music and historical knowledge.

Error Analysis GPT-4o correctly answered 241 questions that Claude 3.5 Sonnet failed to answer, many of which belonged to the hard category, including 109 from History and 48 from Drama and Theatre. Qwen2.5 and LLaMA-3.1 answered 220 and 190 of these missed questions, respectively. These findings suggest that while closed models such as GPT-4o and Claude perform reasonably well in understanding general question structure and intent, they often fall short in capturing localized cultural nuances. This suggests the necessity of developing models that are better attuned to the cultural and linguistic context of Sinhala.

9 Conclusion

We introduce SinhalaMMLU, the first multi-task language understanding benchmark designed to evaluate LLM capabilities in Sinhala, a low-resource language. This benchmark provides a structured evaluation based on the Sri Lankan educational curriculum. While closed-source models demonstrate higher levels of accuracy overall, both open and closed models struggle to achieve comparable results on culturally embedded concepts and language-specific features. This paved the need for future research on culturally aware and low-resource LLMs. We believe that SinhalaMMLU will empower researchers to effectively evaluate and design Sinhala LLMs.

10 Limitations

Although we believe SinhalaMMLU will significantly contribute to Sinhala NLP and the design of next multilingual LLMs, it does have some limitations.

Text-based question only SinhalaMMLU is limited to text-based multiple-choice questions, enabling standardized evaluation but excluding multimodal and open-ended tasks. Extending the benchmark to include images, tables, diagrams, audio and essay-style questions remains a valuable direction for future work.

Mathematics Not Included We specifically exclude mathematics questions because the original exam papers do not include multiple-choice formats for this subject, focusing instead on problem-solving. Additionally, Maths questions are already well covered by existing English math reasoning benchmarks.

Educational Scope The dataset primarily targets the secondary and collegiate levels of the Sri Lankan education system. Expanding the benchmark to include primary-level content, professional-level content, open-ended questions, and tasks that assess generative capabilities would provide a more comprehensive evaluation of LLMs in Sinhala.

Human evaluation Human evaluation was not conducted in this study due to practical limitations. Evaluating subject-specific questions requires age-appropriate annotators with domain expertise, ideally during exam periods, which is logistically challenging. As an alternative, we provide official O-Level and A-Level pass rates as a reference[7]. However, we note that these statistics are drawn from official results for Sri Lanka's G.C.E. Ordinary Level and Advanced Level exams, and reflect only the overall pass rates of student cohorts for each year. While they are not directly comparable to our benchmark, since we focus solely on multiple-choice questions whereas the national exams include structured and essay-type questions, they still provide a meaningful reference point for understanding model performance in real-world educational settings.

Domain Imbalance and Sampling Bias Due to the curriculum structure and limited access to Sinhala exam papers with marking schemes, subjects in domains like humanities and social sciences are overrepresented. This imbalance may affect subject diversity and model performance comparisons.

Prompting Strategies and Evaluation Framework We did not experiment with chain-of-thought prompting or vision-language models, which may yield different results compared to our current approach. While lm-evaluation-harness (Gao et al., 2024) is used for evaluation recently, we adopted a custom evaluation pipeline aligned with prior work (Poh et al., 2024; Koto et al., 2024; Li et al., 2024).

11 Ethical Statement

The SinhalaMMLU dataset used in our study is collected from publicly available web resources, which are published by the government. In compliance with the Sri Lankan Copyright Act Number 36 year 2003[8], specifically section 11, the fair use of a work, including such use by reproduction in copies or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship or research, shall not be an infringement of copyright. Our dataset is intended solely for research and educational purposes and complies with these fair use provisions. The SinhalaMMLU benchmark will be released under an appropriate open license to support research and educational use.

We are committed to ensuring that the data collection and evaluation processes for SinhalaMMLU are conducted with the highest standards of transparency and fairness. To support this, we adopted a crowd-sourcing approach for the annotation process, welcoming contributions from volunteers for data collection. Individuals who provide significant contributions will be offered co-authorship, in accordance with the authorship guidelines set by the ACL, available at Authorship Changes Policy for ACL Conference Papers. The annotators who contributed to the human evaluation will be acknowledged in the acknowledgments section after the review process, in accordance with anonymity guidelines.

Note that in this work, we used an AI assistant tool, Copilot, for coding support. Additionally, our dataset may not fully reflect real-world exams, which often include multimodal and essay-style questions. This limitation should be considered when generalizing our findings to broader educational or practical applications.

[7] https://www.gazette.lk/2025/05/gce-ordinary-level-in-sri-lanka.html

[8] https://www.gov.lk/elaws/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf

Acknowledgments

We would like to thank our annotators, Kavindu Akalanka Wickramasinghe and Dhanushi nagodage, for their valuable contributions to the human evaluation of STEM subjects.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. 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 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Anthropic. 2024. Introducing the next generation of Claude. Anthropic News. Accessed: 2025-05-04.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and

Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Omid Ghahroodi, Marzia Nouri, Mohammad V. Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *ArXiv*, abs/2404.06644.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

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

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

Kushan Hewapathirana, Nisansa de Silva, and C.D. Athuraliya. 2024. M2ds: Multilingual dataset for¬†multi-document summarisation. In *Advances in Computational Collective Intelligence*, pages 219–231, Cham. Springer Nature Switzerland.

Meng Ji, Meng Ji, Pierrette Bouillon, and Mark Seligman. 2023. *Cultural and Linguistic Bias of Neural Machine Translation Technology*, page 100–128. Studies in Natural Language Processing. Cambridge University Press.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

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, and 5 others. 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.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Yang Liu, Maosong Sun, and Erhong Yang. 2024. Omgeval: An open multilingual generative evaluation benchmark for large language models. *CoRR*, abs/2402.13524.

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, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Soon Chang Poh, Sze Jue Yang, Jeraelyn Ming Li Tan, Lawrence Leroy Tze Yao Chieng, Jia Xuan Tan, Zhenyu Yu, Foong Chee Mun, and Chee Seng Chan. 2024. MalayMMLU: A multitask benchmark for the low-resource Malay language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 650–669, Miami, Florida, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

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.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. 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 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Disura Warusawithana, Nilmani Kulaweera, Lakshan Weerasinghe, and Buddhika Karunarathne. 2022. A systematic approach to derive a refined speech corpus for Sinhala. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5107–5113, Marseille, France. European Language Resources Association.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. 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 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

A Data Distribution

In this section, we provide detailed descriptions of the subjects categorized under each domain in Table 2, and present the number of questions per subject across difficulty levels in Table 10. Questions for the "Easy" and "Medium" levels were collected from official examination papers with answer schemes, published between 2017 and 2023. While most "Hard" level questions were sourced from 2012 to 2023, in some subjects it was necessary to extend the search to the early 2000s in order to obtain appropriate questions and their corresponding answer schemes.

No.	Subject	Easy	Medium	Hard
1	History	154	121	-
2	Drama and Theatre	131	140	143
3	Dancing	111	171	-
4	Eastern Music	146	113	-
5	Arts	103	133	-
6	Buddhism	165	120	147
7	Catholicism	135	119	-
8	Christianity	133	118	150
9	Islam	111	120	119
10	Citizenship Education	126	156	-
11	Health and Physical Science	133	105	-
12	Geography	120	157	127
13	Science	167	112	-
14	Sinhala Language and literature	158	120	119
15	Business and Accounting Studies	-	122	109
16	Entrepreneurship Studies	-	117	-
17	Home Economics	-	150	140
18	Communication and Media Studies	-	119	112
19	Agriculture and Food Technology	-	136	158
20	Design and Construction Technology	-	117	-
21	Economics	-	-	129
22	Biosystems Technology	-	-	106
23	Buddhist Civilization	-	-	119
24	Political Science	-	-	160
25	Physics	-	-	107
26	Chemistry	-	-	116
27	Biology	-	-	127
28	Oriental Music	-	-	117
29	History of Sri Lanka	-	-	147
30	Dancing Indigenous	-	-	132

Table 10: Subject-wise counts of questions by each difficulty level in the SinhalaMMLU dataset.

Educational Levels Table 11 summarizes the educational levels in Sri Lanka.

A.1 Examples

Figure 6 illustrates example questions from the remaining domains, Business and Other, while Figure 7 depicts a negation example. Examples of culturally relevant questions are provided in Figure 8.

Level	Grades	Age Range	Description
Primary (Excluded)	1–5	5–9	Focuses on foundational learning (e.g., drawing, writing, filling blanks); excluded from SinhalaMMLU due to unsuitability for MCQs. Ends with the optional Grade 5 Scholarship Exam.
Junior Secondary	6–9	10–13	Covers core subjects; first phase of secondary education. Includes nine subjects as first language, English, mathematics, science and technology, social studies, life skills, religion, aesthetics, health, and physical education. A second language is also introduced.
Senior Secondary (O-Level)	10–11	14–16	Prepares students for GCE O-Level exams with six main and three additional subjects.
Collegiate (A-Level)	12–13	17–19/20	Includes five academic streams: Arts, Commerce, Biological Science, Physical Science, and Technology. Students select three core subjects + General English + Common General Test. Qualification for university admission.

Table 11: Structure of the Sri Lankan national education curriculum relevant to SinhalaMMLU.

Business studies : Entrepreneurship studies O/L

වර්තමානයේ ඇමරිකානු ඩොලරයට සාපේක්ෂව ශ්‍රී ලංකා රුපියලෙහි හුවමාරු අගය වෙනස් වීම අදාළ වන්නේ
1. ගෝලීය පරිසරයට ය.
2. ප්‍රජා පරිසරයට ය.
3. දේශපාලන හා නෛතික පරිසරයට ය.
4. ආර්ථික පරිසරයට ය.

The current exchange rate of the Sri Lankan Rupee against the US Dollar is related to
1.. the global environment.
2. the societal environment.
3. the political and legal environment.
4. the economic environment.

Other : Agricuture O/L

නියමිත මේරීමේ අවස්ථාවේ දී අස්වනු නෙළීමට ප්‍රමාද වුවහොත් තන්තුමය ස්වභාවය වැඩි වීම නිසා පරිභෝජනයට ගත නොහැකි වන බෝග අස්වනු මොනවා ද?
1. මුං හා කව්පී
2. අඹ හා තක්කාලි
3. බණ්ඩක්කා හා වැටකොළු
4. පැපොල් හා කෙසෙල්.

What are the crops that cannot be consumed due to the increase in fibrous nature if they are not harvested at the proper maturity?
1. Green gram and cowpea
2. Mango and tomato
3. Okra and betel nut
4. Papaya and banana.

Figure 6: Example questions from the Business and Other domains.

Social Science : Political Science - Negation question

පහත ප්‍රශ්නයේ මාතෘකාව හා සම්බන්ධ ප්‍රකාශ පහක් ඉදිරිපත් කොට ඇත. ඒවායින් එක් ප්‍රකාශයක් **සාවද්‍ය වේ**. සාවද්‍ය ප්‍රකාශය තෝරන්න.
නූතන ප්‍රජාතන්ත්‍රවාදී ආණ්ඩුක්‍රම තුළ ව්‍යවස්ථාදායකය:

1. සාමාන්‍යයෙන් සමන්විත වන්නේ ජනතාව තෝරාපත් කරන ගිහි දේශපාලනඥයන්ගෙනි.
2. සමස්ත ජාතියම නියෝජනය කරන නියෝජිතයින්ගෙන් සමන්විත විශාල මණ්ඩලයක් වේ.
3. විවාද කිරීමේ ප්‍රධාන ජාතික සභාව ලෙස සැලකේ.
4. ආණ්ඩුවේ කාර්යයන්ට සුජාත භාවය සැපයීමේ වගකීම දරනු ලැබේ.
5. ජාතික අයවැය සම්පාදනය කිරීමේ හා ක්‍රියාවට නැංවීමේ වගකීම දරනු ලැබේ.

Five statements are given related to the topic of the question below. One of them is false. Choose **the false statement**.
In modern democratic systems, the legislature:

1. Usually consists of lay politicians elected by the people.
2. Is a large body of representatives representing the entire nation.
3. Is considered the main debating body of the nation.
4. Is responsible for providing legitimacy to the work of the government.
5. Is responsible for preparing and implementing the national budget.

Figure 7: Example of negation question.

B Experimental Setup

All model generations were conducted on a single NVIDIA RTX A6000 Ada GPU. To optimize memory efficiency, we applied 4-bit quantization during inference. Experiments with GPT-4o were run using top_p = 0 and temperature = 0.5.

Model details The open-source model and closed model artifacts used in our experiments are listed in Table 13 and are accessible via the Hugging Face Hub except the GPT-4o and Claude.

අතීත ශ්‍රී ලංකාවේ අභ්‍යන්තර වෙළෙඳාම පිළිබඳ නීති ඇතුළත් ශිලා ලේඛනයක් වන්නේ,
1. ගොඩවාය සෙල්ලිපියයි.
2. මිහින්තලා පුවරු ලිපියයි
3. වේවැල්කැටිය සෙල්ලිපියයි
4. සොරබොර වැව ටැම් ලිපියයි

An inscription in which the rules governing internal trade in ancient Sri Lanka were included is
1 Sorabora wewa inscription.
2. Mihintala slab inscription.
3. Wewelketiya inscription.
4. Godawaya inscription.

පුරාණ කමත් භාෂාවේ ' අළුහන් වැඩීම' යනු
1. කොළ මැඩීමට හරකුන් එකට ඇඳා ගැනීම යි.
2. අලුයම් කාලයේ කමතට ගොයම් ගෙන ඒම යි.
3. අළු වලින් කමතේ යන්ත්‍රයක් ඇඳීම යි.
4. ගොයම් පෑගූ පසු පිදුරු ඉවත් කිරීම යි.

In ancient Kamath language, 'Aluhan Wadima' means
1. Tying oxen together to trample the threshing floor
2. Bringing paddy (rice) to the threshing floor in the early morning.
3. Drawing a ritual design/yantra on the threshing floor with ashes
4. Removing straw after the paddy has been threshed.

මහනුවර බිතු සිතුවම් කලාවට ආවේණික ලක්ෂණයක් නොවන්නේ ,
1 . තීරු චිත්‍ර සංරචන ක්‍රමය
2 . ද්විමාන ලක්ෂණ සහිත පැතලි වර්ණ භාවිතය
3 . ත්‍රිමාණ ලක්ෂණ සහිත පර්යාවලෝකනය
4 . පාර්ශවදර්ශී ලෙස රූප නිර්මාණය

Which is NOT a characteristic feature of Kandyan mural art?
1. Strip/band composition method
2. Two-dimensional features with flat colors
3. Three-dimensional features with perspective
4. Creation of figures in profile view

Figure 8: Examples of cultural questions from history, language, and arts.

B.1 Prompt

For the SinhalaMMLU evaluation, we used two distinct prompt templates, one that specifies the subject domain and another that omits this information. Both follow the instruction format proposed by Hendrycks et al. (2021), with prompt instructions given in the same language as the question. We observed that including the subject name generally improves performance across models. Therefore, for all subsequent experiments, we report results using prompts that include subject information. For evaluation, we use the following prompting formats in 9. For 3-shot evaluation, the identical format is repeated. The fewshot examples were gathered from random questions for each subject. For closed models, we additionally use a system prompt to specify the expected answer format. All experiments involving closed models were conducted using the official APIs provided by OpenAI and Anthropic.

With subject information

මෙය **[SUBJECT]** විෂයයට අදාළ බහුවරණ ප්‍රශ්නයකි. පහත ප්‍රශ්නයට 1, 2, 3, 4 යන පිළිතුරුවලින් නිවැරදි හෝ ඉතාමත් ගැළපෙන හෝ පිළිතුර තෝරන්න.

ප්‍රශ්නය: [Question]
[OPTIONS]
පිළිතුර:

```
This is a multiple-choice question related to the [SUBJECT].
Choose the correct or most appropriate answer from answers
1, 2, 3, or 4 for the following question.

Question: [Question]
[OPTIONS]
Answer:
```

With out subject information

මෙය බහුවරණ ප්‍රශ්නයකි. පහත ප්‍රශ්නයට 1, 2, 3, 4 යන පිළිතුරුවලින් නිවැරදි හෝ ඉතාමත් ගැළපෙන හෝ පිළිතුර තෝරන්න.

ප්‍රශ්නය: [Question]
[OPTIONS]
පිළිතුර:

```
This is a multiple-choice question. Choose the correct or
most appropriate answer from answers 1, 2, 3, or 4 for the
following question.

Question: [Question]
[OPTIONS]
Answer:
```

Figure 9: Prompt template used in the task.

C Additional Results and Discussions

Does including the subject affect performance? To evaluate the impact of the subject, we compared two prompt variants: with and without the subject. As shown in Figure 11, Claude 3.5 Sonnet's accuracy improves across all domains when the subject is included, while GPT-4o performs worse on Sinhala language tasks. This suggests GPT may have weaker prior knowledge of Sinhala grammar or context compared to Claude, making it more reliant on explicit prompting.

C.1 Results by difficulty

Table 12 presents the detailed zero-shot results across six domains for each difficulty level.

C.2 Results By Subject

Subject-wise analysis shows that models tend to achieve their highest scores in subjects that require less analytical reasoning, such as Buddhism, Christianity, Islam Citizenship Education, and Health, where questions are more fact-based and rely on descriptive knowledge. In contrast, lower scores are observed in culturally grounded subjects like Sinhala language, History, and Drama and Theatre, which require deeper cultural and contextual understanding. A more detailed breakdown of subject-

Figure 10: Performance (% accuracy) breakdown across the 30 subjects. “E”, “M”, and “H” indicate each difficulty level: easy, medium, and hard, respectively.

Model	Easy					Medium							Hard						
	Hum	Lang	SocSci	STEM	Avg	Hum	Lang	SocSci	STEM	Bus	Other	Avg	Hum	Lang	SocSci	STEM	Bus	Other	Avg
claude-3-5-sonnet	71.69	68.39	82.43	78.66	74.18	68.09	70.94	76.77	79.82	74.25	66.67	70.43	57.94	45.69	72.24	47.21	71.98	64.48	60.12
claude-3-haiku	45.87	46.45	62.16	50.00	49.54	42.46	48.72	52.08	44.95	50.64	39.80	44.65	34.06	35.34	33.10	25.22	37.50	39.48	34.22
gpt-4o-2024-08-06	68.42	50.32	74.59	77.44	68.94	62.59	58.97	71.39	78.90	67.38	62.55	65.00	54.33	44.83	67.26	46.63	67.67	60.32	56.71
Qwen2.5-7B	26.68	25.16	25.68	18.90	25.66	22.16	20.51	27.14	23.85	24.46	23.33	23.42	17.51	18.10	23.49	16.72	15.95	20.04	18.46
Qwen2.5-7B-chat	30.03	28.39	35.41	28.66	30.85	28.99	23.93	27.63	35.78	27.04	27.06	28.25	24.74	21.55	19.93	26.10	27.16	20.24	23.56
Qwen2.5-32B	33.39	35.48	38.11	34.15	34.58	29.96	23.08	36.67	30.28	30.04	30.20	30.81	22.65	24.14	29.54	21.41	20.26	24.80	23.52
Qwen2.5-32B-chat	41.39	38.71	43.51	49.39	42.30	37.06	33.33	38.88	47.71	35.19	39.02	37.87	31.11	24.14	33.10	30.50	34.48	28.97	30.81
Qwen2.5-72B	40.71	33.55	42.97	46.34	41.06	35.64	32.48	44.50	51.38	32.62	36.67	37.55	27.02	18.97	31.67	29.62	25.00	32.34	28.40
Qwen2.5-72B-chat	43.29	38.06	51.08	51.22	45.11	41.40	47.86	48.41	60.55	45.06	40.20	43.77	34.35	31.03	35.23	38.12	37.50	37.50	35.72
Llama-3-8B	26.94	25.16	26.22	22.56	26.26	22.34	21.37	27.87	24.77	24.89	23.53	23.78	17.70	17.24	24.20	16.42	16.81	19.44	18.50
Llama-3-8B-chat	27.97	24.52	27.84	22.56	27.17	22.34	20.51	28.36	22.94	23.61	24.51	23.82	17.79	18.97	23.84	16.72	15.95	21.83	19.01
Llama-3-70B	26.25	27.10	24.59	29.88	26.31	22.61	24.79	27.38	25.69	23.61	24.31	24.06	17.70	17.24	22.78	17.89	15.95	19.84	18.53
Llama-3-70B-chat	29.95	23.23	28.92	28.66	29.07	28.28	22.22	25.92	25.69	20.60	26.86	26.50	22.84	14.66	23.49	20.82	20.69	18.06	21.11
Llama-3.1-8B	26.33	25.16	28.92	20.73	26.26	22.78	26.50	29.83	26.61	27.47	24.12	24.98	19.22	18.10	21.00	16.72	19.40	22.62	19.72
Llama-3.1-8B-chat	28.83	29.68	34.05	23.17	29.44	26.68	21.37	31.05	22.02	24.46	26.86	26.78	19.89	24.14	19.57	21.99	17.67	23.02	20.75
Llama-3.1-70B	26.16	26.45	26.76	23.78	26.09	21.99	23.08	28.36	21.10	23.61	23.33	23.46	17.60	18.10	23.13	17.01	15.95	19.84	18.46
Llama-3.1-70B-chat	30.12	30.97	32.16	28.05	30.42	27.13	27.35	24.45	21.10	33.48	29.02	27.41	24.17	20.69	19.93	22.58	19.83	21.43	22.38
Llama-3.2-1B	26.59	25.16	25.41	18.90	25.55	22.16	20.51	27.14	22.94	23.61	23.53	23.34	17.32	18.10	23.84	16.72	15.95	19.84	18.38
Llama-3.2-1B-chat	26.76	25.16	25.68	18.90	25.72	22.25	20.51	27.14	22.94	23.61	23.53	23.38	17.32	18.10	23.84	17.30	15.95	19.84	18.46
Llama-3.2-3B	26.59	27.74	25.68	18.90	25.82	22.16	20.51	26.89	22.94	23.18	23.53	23.26	17.32	18.10	23.49	16.72	15.95	19.84	18.34
Llama-3.2-3B-chat	26.68	24.52	25.68	19.51	25.66	22.25	21.37	26.89	22.02	22.75	23.53	23.26	17.41	18.10	23.84	17.01	15.95	19.84	18.46
Llama-3.3-70B-chat	28.83	26.45	26.76	29.27	28.25	23.85	21.37	30.32	22.94	29.18	25.69	25.62	20.55	21.55	23.49	20.23	22.41	20.04	20.95
Mistral-7B	26.59	25.16	25.41	18.90	25.55	22.16	20.51	27.14	22.94	23.61	23.53	23.34	17.32	18.10	23.84	16.72	15.95	18.91	18.14
Mistral-7B-chat	23.92	22.58	25.91	27.44	24.49	22.96	27.35	32.27	22.94	23.18	21.76	24.46	17.41	18.97	20.28	18.42	16.81	19.84	18.37
aya-expanse-8b	24.87	23.87	26.76	21.34	24.85	23.32	29.91	25.67	26.61	23.18	23.33	24.14	19.89	18.97	22.06	16.72	17.67	20.04	19.49
aya-expanse-32b	22.20	21.94	35.14	28.05	25.28	27.39	33.33	28.61	33.03	26.61	26.67	27.89	22.26	20.69	25.27	21.11	20.69	22.62	22.30

Table 12: Zero-shot model accuracy across six domains for each difficulty level (easy, medium, and hard.

Model	Source
GPT-4o	`gpt-4o-2024-08-06`
claude 3.5 sonnet	`claude-3-5-sonnet-20241022`
claude-3-haiku	`claude-3-haiku-20240307`
aya-expanse-32b	`CohereForAI/aya-32b`
aya-expanse-8b	`CohereForAI/aya-8b`
Llama-3.2-1B-chat	`meta-llama/Llama-3-1B-instruct`
Llama-3.2-3B-chat	`meta-llama/Llama-3-3B-instruct`
Llama-3.3-70B-chat	`meta-llama/Llama-3-70B-instruct`
Llama-3.2-1B	`meta-llama/Llama-3-1B`
Llama-3.2-3B	`meta-llama/Llama-3-3B`
Llama-3-70B-chat	`meta-llama/Meta-Llama-3-70B-Instruct`
Llama-3-8B-chat	`meta-llama/Meta-Llama-3-8B-Instruct`
Llama-3.1-70B-chat	`meta-llama/Meta-Llama-3.1-70B-Instruct`
Llama-3.1-8B-chat	`meta-llama/Meta-Llama-3.1-8B-Instruct`
Llama-3-70B	`meta-llama/Meta-Llama-3-70B`
Llama-3-8B	`meta-llama/Meta-Llama-3-8B`
Llama-3.1-70B	`meta-llama/Meta-Llama-3.1-70B`
Llama-3.1-8B	`meta-llama/Meta-Llama-3.1-8B`
Mistral-7B-v0.3	`mistralai/Mistral-7B-v0.3`
Mistral-7B-chat-v0.3	`mistralai/Mistral-7B-Instruct-v0.3`
Qwen2.5-32B	`Qwen/Qwen2.5-32B`
Qwen2.5-32B-chat	`Qwen/Qwen2.5-32B-Instruct`
Qwen2.5-72B	`Qwen/Qwen2.5-72B`
Qwen2.5-72B-chat	`Qwen/Qwen2.5-72B-Instruct`
Qwen2.5-7B	`Qwen/Qwen2.5-7B`
Qwen2.5-7B-chat	`Qwen/Qwen2.5-7B-Instruct`

Table 13: Lists of the LLMs we used in this study and their corresponding Hugging Face IDs except for GPT-4o and Claude.

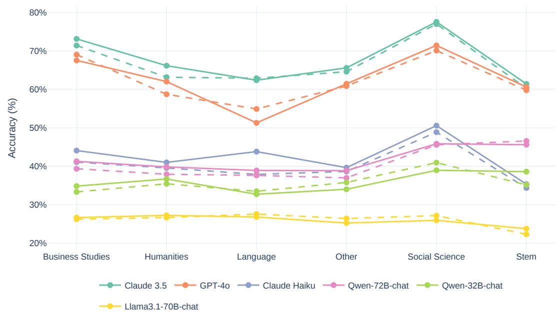

Figure 11: Accuracy variation of LLMs across 6 domains under two prompting strategies with and without including the subject in the prompt. Solid lines indicate prompts including the subject; dashed lines indicate prompts without it. Each model is represented by a consistent color.

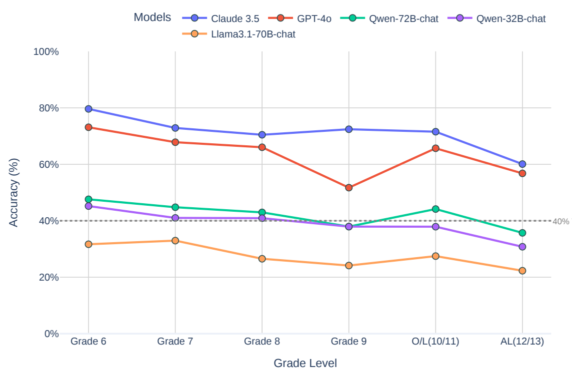

Figure 12: LLM accuracy by grade level with a 40% threshold line. Grades with no questions are excluded.

wise performance is provided in the Figure 10.

C.3 Results by grade

As illustrated in Figure 12, SinhalaMMLU includes metadata on educational grade levels, enabling a more fine-grained evaluation of LLM performance from an educational perspective. In the Sri Lankan education system, a minimum score of 40% is typically required to pass. However, the current benchmark evaluates only the multiple-choice component of examination papers, which represent 40–50% of the total marks, the remainder consisting of structured and essay-type questions. Interestingly, performance drops at grades 9, with models performing better on GCE O-Level exam questions, this is basically due to that grade 9 only contained questions in arts which skewed the results. Most open source models barely reach this threshold, suggesting limited readiness for real exam scenarios.

C.4 Few-shot vs zero-shot

Table 14 presents the detailed few-shot results across six domains, while Table 15 compares few-shot and zero-shot performance for each difficulty level.

Model	Humanities	Language	Social Science	Stem	Business Studies	Other	Average
LLAMA-3-8B	24.84	26.55	30.19	23.55	20.65	21.33	24.92
LLAMA-3-8B-CHAT	22.45	19.59	25.44	19.22	19.78	20.77	22.03
LLAMA-3-70B	22.61	21.39	25.40	22.16	20.65	21.77	22.59
LLAMA-3-70B-CHAT	23.30	23.16	24.05	27.20	19.14	23.51	23.35
LLAMA-3.1-8B	28.28	27.32	34.43	27.15	26.88	25.19	28.60
LLAMA-3.1-8B-CHAT	23.47	23.71	27.90	23.29	21.94	23.20	23.99
LLAMA-3.1-70B	23.29	22.68	25.19	21.16	19.78	21.44	22.90
LLAMA-3.1-70B-CHAT	25.16	26.47	29.15	23.43	20.86	23.20	25.05
LLAMA-3.2-1B	23.29	20.10	25.09	22.16	23.87	22.98	23.30
LLAMA-3.2-1B-CHAT	22.18	21.65	25.64	18.40	19.78	22.21	22.17
LLAMA-3.2-3B	23.86	20.88	25.00	21.76	22.80	22.65	23.47
LLAMA-3.2-3B-CHAT	22.69	21.39	25.05	21.17	20.22	19.89	22.28
LLAMA-3.3-70B-CHAT	26.09	24.63	30.31	24.18	20.00	21.74	25.45
MISTRAL-7B-CHAT	22.87	22.68	25.83	21.56	24.09	21.66	23.13
QWEN2.5-32B	28.94	28.87	32.71	30.62	25.16	31.05	29.68
QWEN2.5-32B-CHAT	37.29	35.05	41.45	39.74	32.47	36.24	37.54
QWEN2.5-72B	35.05	27.58	43.87	36.16	29.89	30.94	35.20
QWEN2.5-72B-CHAT	40.59	36.86	46.70	44.30	39.57	37.90	41.24
QWEN2.5-7B	22.18	21.65	25.64	18.40	19.78	22.10	22.15
QWEN2.5-7B-CHAT	28.23	26.55	31.04	26.06	25.38	28.40	28.18
AYA-EXPANSE-32B	24.91	25.74	30.26	27.20	21.72	22.40	25.21
AYA-EXPANSE-8B	22.30	21.65	25.64	17.43	20.22	21.77	22.11

Table 14: Few-shot accuracy per each domain.

Model	Easy		Medium		Hard	
	Zero-Shot	3-Shot	Zero-Shot	3-Shot	Zero-Shot	3-Shot
LLAMA-3.2-1B	25.55	25.77	23.34	24.74	18.38	19.76
LLAMA-3.2-1B-CHAT	25.72	25.54	23.38	23.34	18.46	18.42
LLAMA-3.2-3B	25.82	26.63	23.26	24.50	18.34	19.80
LLAMA-3.2-3B-CHAT	25.66	26.76	23.26	22.98	18.46	18.21
LLAMA-3.3-70B-CHAT	28.25	29.63	25.62	25.14	20.95	20.64
LLAMA-3-70B	26.31	27.11	24.06	24.98	18.53	18.41
LLAMA-3-70B-CHAT	29.07	25.26	26.50	25.36	21.11	19.47
LLAMA-3-8B	26.26	29.66	23.78	26.50	18.50	19.41
LLAMA-3-8B-CHAT	27.17	24.38	23.82	23.46	19.01	18.79
LLAMA-3.1-70B	26.09	25.72	23.46	24.66	18.46	18.71
LLAMA-3.1-70B-CHAT	30.42	28.80	27.41	26.14	22.38	19.81
LLAMA-3.1-8B	26.26	34.31	24.98	30.65	19.72	21.80
LLAMA-3.1-8B-CHAT	29.44	28.08	26.78	25.14	20.75	19.74
MISTRAL-7B-CHAT	24.49	25.15	24.46	25.46	18.37	19.02
QWEN2.5-32B	34.58	36.26	30.81	27.77	23.52	26.74
QWEN2.5-32B-CHAT	42.30	41.57	37.87	40.42	30.81	31.54
QWEN2.5-72B	41.06	49.05	37.55	25.54	28.40	34.60
QWEN2.5-72B-CHAT	45.11	46.25	43.77	43.85	35.72	34.69
QWEN2.5-7B	25.66	25.54	23.42	23.30	18.46	18.42
QWEN2.5-7B-CHAT	30.85	31.01	28.25	30.41	23.56	23.76
AYA-EXPANSE-32B	25.28	28.75	27.89	25.54	22.30	19.60
AYA-EXPANSE-8B	24.85	25.59	24.14	23.50	19.49	18.05

Table 15: Model Performance Comparison (Zero-Shot vs 3-Shot).