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

Soon Chang Poh, Sze Jue Yang, Jeraelyn Ming Li Tan, Lawrence Leroy Tze Yao Chieng, Jia Xuan Tan, Zhenyu Yu, Chee Mun Foong, Chee Seng Chan

UNIVERSITI MALAYA

SINCE 1955

EMNLP 2024

Introduction

Malay is low resource language

- Malay language's presence on the web content is only 0.1%, 500 times smaller than English, and 10 times smaller than Indonesian and Vietnamese.

Language	Web content (%)
English	50.0
Indonesian, Vietnamese	1.0
Malay	**0.1**

Table 1. Language distribution of web content

- Existing benchmark (Seabench) and the training datasets for Southeast Asian Large Language Models (LLMs) like SeaLLM and Sailor include only a limited amount of content in Malay language.

Category	Name	Amount of Malay content
Benchmark	Seabench	<100 questions
LLMs' training dataset	SeaLLM	<2% of data
	Sailor	<5% of data

Table 2. Amount of Malay content in existing benchmark and training dataset

- The performance of LLMs and Large Vision-Language Models (LLVMs) on Malay remains under-explored.

Malay is vital in Malaysia.

- As the national language of Malaysia with 30 million speakers, it is widely used in government communications, legal documents, media, and public signage.

MalayMMLU

- Curated based on Malaysia's standard education curriculum
- Contains 22 subjects in 5 topics in primary and secondary education level

Level	Topic	# Questions	%
Primary (33.6%)	Language	4,684	19.3
	Humanities	1,721	7.1
	STEM	2,24	0.9
	Social Science	1,078	4.5
	Others	426	1.8
Secondary (66.4%)	Language	1,604	6.6
	Humanities	2,674	11.0
	STEM	2,219	9.2
	Social Science	5,840	24.1
	Others	3,743	15.5
Total		**24,213**	**100.0**

Table 3: Data distribution by education level and topic

- Social science topic (secondary) has highest number of questions.
- Questions for secondary schools are longer in average.

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

Table 4. Average question and answer length (in characters)

Sample

- MalayMMLU contains questions about local context such as history and culture (Fig. 1).
- Original cultural sense meaning of idioms are lost due to the imprecise English translations (Fig. 2).

History (Form 1)	
Mengapakah orang laut di Melaka sangat penting semasa pemerintahan Parameswara? **A. Menjaga keselamatan laut Melaka** B. Menangkap Ikan	Why were seafarers in Malacca so important during the reign of Parameswara? **A. Maintaining the safety of Malacca's sea** B. Catching Fish
History (Standard 6)	
Tarian zapin merupakan satu warisan seni negara **A. Betul** B. Salah	Zapin dance is a national art heritage **A. That's right** B. Wrong

Figure 1: Questions with Malaysian context (Left) original, (right) Google translation in English

Malay (Form 3)	
Pilih peribahasa yang sesuai berdasarkan situasi yang diberikan. Duit raya yang diterima oleh kanak-kanak wajar dimanfaatkan sebaik-baiknya dengan cara menyimpannya di dalam bank untuk masa depan mereka. Amalan menabung merupakan satu tindakan yang baik dan mengajar seseorang berjimat cermat. A. Bertanam tebu di tepi bibir **B. Sikit-sikit lama-lama jadi bukit** C. Bagai belut pulang ke lumpur	Choose the appropriate proverb based on the given situation. Raya money received by children should be used as best as possible by keeping it in the bank for their future. The practice of saving is a good action and teaches a person to be thrifty. A. Planting sugar cane on the edge of the lip **B. Little by little it becomes a hill** C. Like an eel returning to the mud

Figure 2: A question about Malay idioms. (Left) original, (right) Google translation in English

Key Results

Organization	Model	Vision	Language	Humanities	STEM	Social Science	Others	Average
	Random		38.01	42.09	36.31	36.01	38.07	38.02
OpenAI	**GPT-4o**	✓	**87.12**	**88.12**	**83.83**	**82.58**	**83.09**	**84.98**
	GPT-4	✓	82.90	83.91	78.80	77.29	77.33	80.11
	GPT-4o mini	✓	82.03	81.50	78.51	75.67	76.30	78.78
	GPT-3.5		69.62	71.01	67.17	66.70	63.73	67.78
Meta	LLaMA-3.1 (70B)		78.75	82.59	78.96	77.20	75.32	78.44
	LLaMA-3.1 (8B)		65.47	67.17	64.10	62.59	62.13	64.24
	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
	LLaMA-3.2 (3B)		58.52	60.66	56.65	54.06	52.75	56.45
	LLaMA-3.2 (1B)		38.88	43.30	40.65	40.56	39.55	40.46
Qwen (Alibaba)	Qwen 2.5 (72B)		79.09	79.95	80.88	75.80	75.05	77.79
	Qwen-2.5 (32B)		76.96	76.70	79.74	72.35	70.88	74.83
	Qwen-2-VL (7B)	✓	68.16	63.62	67.58	60.38	59.08	63.49
	Qwen-2-VL (2B)	✓	58.22	55.56	57.51	53.67	55.10	55.83
	Qwen-1.5 (7B)		64.47	60.64	61.97	57.66	58.05	60.47
	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
Zhipu	GLM-4-Plus		78.04	75.63	77.49	74.07	72.66	75.48
	GLM-4-Air		67.88	69.56	70.20	66.06	66.18	67.60
	GLM-4-Flash		63.52	65.69	66.31	63.21	63.59	64.12
	GLM-4		63.39	56.72	54.40	57.24	55.00	58.07
Google	Gemma-2 (9B)		75.83	72.83	75.07	69.72	70.33	72.51
	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
SAIL (Sea)	Sailor (14B)		78.40	72.88	69.63	69.47	68.67	72.29
	Sailor (7B)		74.54	68.62	62.79	64.69	63.61	67.58
Damo (Alibaba)	SeaLLM-v2.5 (7B)		69.75	67.94	65.29	62.66	63.61	65.89
Mistral	Pixtral (12B)		64.81	62.68	64.72	63.93	59.49	63.25
	Mistral Small (22B)		65.19	65.03	63.36	61.58	59.99	63.05
	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
Microsoft	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

*Table 5: Zero-shot performance of LLMs/LVLMs (first token accuracy) in MalayMMLU. The highest accuracy is **bolded***

Best Performer

- **Overall** : GPT-4o
- **Open-source model** : LLaMA-3.1 (70B)
- **<= 50B parameters** : Qwen-2.5 (32B)
- **<= 10B parameters** : Gemma-2 (9B)
- **<= 5B parameters** : LLaMA-3.2 (3B)

LLMs finetuned with SEA datasets

- Southeast Asian (SEA) datasets, such as *Sailor* and *SeaLLMs* (both of which are finetuned from Qwen-1.5 and Gemma, respectively) exhibit enhanced performance.