

SeaExam and SeaBench: Benchmarking LLMs with Local Multilingual Questions in Southeast Asia

Chaoqun Liu^{*12} Wenxuan Zhang^{†23} Jiahao Ying²⁴ Mahani Aljunied²³
Anh Tuan Luu¹ Lidong Bing²³

¹Nanyang Technological University, Singapore; ²DAMO Academy, Alibaba Group, Singapore;
³Hupan Lab, 310023, Hangzhou, China; ⁴Singapore Management University
{chaoqun.liu, saike.zwx}@alibaba-inc.com

Abstract

This study introduces two novel benchmarks, SeaExam and SeaBench, designed to evaluate the capabilities of Large Language Models (LLMs) in Southeast Asian (SEA) application scenarios. Unlike existing multilingual datasets primarily derived from English translations, these benchmarks are constructed based on real-world scenarios from SEA regions. SeaExam draws from regional educational exams to form a comprehensive dataset that encompasses subjects such as local history and literature. In contrast, SeaBench is crafted around multi-turn, open-ended tasks that reflect daily interactions within SEA communities. Our evaluations demonstrate that SeaExam and SeaBench more effectively discern LLM performance on SEA language tasks compared to their translated benchmarks. This highlights the importance of using real-world queries to assess the multilingual capabilities of LLMs.¹

1 Introduction

Large Language Models (LLMs) have shown remarkable performance across various English benchmarks, including both human exam datasets such as MMLU (Hendrycks et al., 2021), or instruction-following datasets such as MT-Bench (Zheng et al., 2023b), indicating their strong capabilities (OpenAI, 2023; Dubey et al., 2024; Team et al., 2024). As these LLMs are increasingly deployed globally, there is growing interest in their ability to handle multiple languages and adapt to a wide range of multilingual applications (Huang et al., 2023; Qin et al., 2024; Huang et al., 2024; Dou et al., 2024; Nguyen et al., 2023; Zhang et al., 2024).

^{*}Chaoqun Liu is under the Joint PhD Program between DAMO Academy and Nanyang Technological University.

[†]Wenxuan Zhang is the corresponding author.

¹SeaExam and SeaBench are publicly available at <https://github.com/DAMO-NLP-SG/SeaExam> and <https://github.com/DAMO-NLP-SG/SeaBench>.

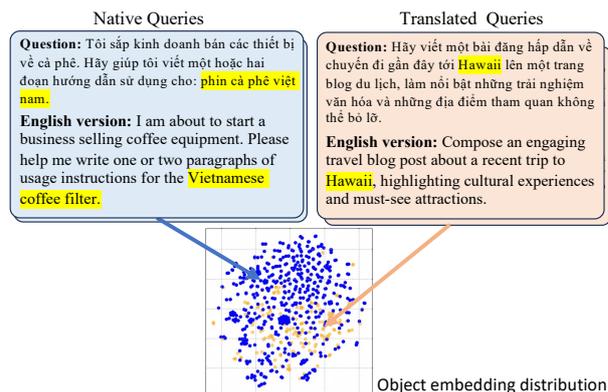


Figure 1: Compared with local usage queries in Vietnamese, questions in English-based translations show more American context (Hawaii). To better illustrate this discrepancy, we extracted the object in these questions and visualised their distribution. The results show that the objects in translated questions cover only a small portion of those in local usage queries.

This led to the development of multiple multilingual benchmarks to assess the multilingual capabilities of LLMs (Lai et al., 2023; Ahuja et al., 2023; Zhang et al., 2023). Among them, many datasets such as MGSM (Shi et al., 2022), XNLI (Conneau et al., 2018), and Multilingual MMLU (Hendrycks et al., 2021; OpenAI, 2023) are typically constructed by translating the English set into target languages. Considering that original English test sets are often carefully designed, such translations provide an effective way to leverage the task categorization, evaluation targets, and construction methods of the monolingual dataset into the multilingual context.

However, such translated questions focus merely on evaluating the same contextual elements as their monolingual counterparts. In other words, they focus primarily on the application scenarios relevant to the original benchmarks rather than adapting to a wide range of multilingual applications in the real world. Instead, a truly effective multilingual benchmark must also consider the content typically

used in the practical application of the target language (Liu et al., 2024). For example, as shown in Figure 1, we visualize the distribution of objects in questions collected from local usage queries versus those translated from English. Compared to local usage queries, translated questions based on English exhibit more of an American context, e.g., involving the place “Hawaii”. It shows that translated questions cover only a small portion of the entities in local usage queries, indicating a significant divergence in the query context.

Considering the scarcity of such effective multilingual benchmarks, this paper introduces two new benchmarks, SeaExam and SeaBench. These benchmarks are specifically designed to address the unique application scenarios and cultural contexts of Southeast Asian (SEA) countries, which often differ significantly from western-centric datasets. Following the design principles of two widely used English-based datasets, MMLU and MT-bench, we do not simply translate the original English questions but incorporate real-world usage scenarios from SEA natives into the content — allowing us to measure a model’s adaptability in multilingual application scenarios. Specifically, SeaExam is a multitask exam dataset sourced from real exams in SEA countries that cover a wide range of subjects including local history, geography, and literature. SeaBench, following MT-Bench’s approach, focuses on multi-turn instruction-following tasks spanning ten task categories. It incorporates scenarios and instructions that are commonly encountered in SEA cultures and daily life.

Our experimental analysis quantitatively demonstrates that, **1)** Compared to the translated benchmarks MMLU and MT-bench, our SeaExam and SeaBench benchmarks include questions that are more aligned with the daily usage of regional languages (Section 3.1). **2)** Furthermore, using SeaExam and SeaBench, we are able to more effectively discern the capabilities of models in real-world multilingual applications (Section 3.2.1). Further analysis reveals that **3)** While multiple-choice questions in exam datasets can objectively measure model capabilities, open-ended questions are more effective in highlighting differences in model performance across various languages (Section 3.2.2 and Section 3.2.3). Additionally, we find that **4)** The nine models involved generally perform poorly in the “safety” category — evaluating whether the models generate harmful responses in the local context (Section 3.2.4). Therefore, we

advocate for enhanced safety measures in multilingual applications to adapt to a broader range of scenarios.

The key contribution can be summarized as:

- We introduce two new benchmarks, SeaExam and SeaBench, which extend the scope of the translated MMLU and MT-bench frameworks to better accommodate the unique linguistic features and practical content contexts of the Southeast Asian (SEA) region.
- We compare these benchmarks with translated counterparts, such as MMLU and MT-Bench, and find that SeaExam and SeaBench have closer distribution to real-world queries. Utilizing these benchmarks allows for a better differentiation of model performance across different language uses.

2 SeaExam and SeaBench

We aim to build multilingual benchmarks to comprehensively evaluate model adaptability to Southeast Asia applications, focusing on both linguistic style and content essence that cannot be fully measured with translated questions. Following the design principle of MMLU and MT-bench, two comprehensive datasets in measuring the English capabilities of large language models, we incorporate real local exams of each country for SeaExam and engage native speakers to craft instructions commonly used in the corresponding language communities for SeaBench. This approach ensures that our benchmarks reflect real-world usage in SEA contexts. We outline the detailed creation processes for SeaExam and SeaBench in Section 2.1 and Section 2.2, respectively.

2.1 SeaExam Construction

Evaluating LLMs using human exam questions can provide valuable insights into the model’s performance, as these questions encompass a wide range of knowledge types. However, relying solely on translations of monolingual exam questions can introduce content biases into model evaluations. For example, the widely used MMLU benchmark includes categories such as “US History”, which may be more relevant to American users.

To address this, we decide to manually collect exam questions from the SEA region (Indonesian (id), Thai (th), and Vietnamese (vi)). We follow the construction of M3Exam (Zhang et al., 2023),

Indonesian (language)	Vietnamese (natural science)	Thai (social science)
<p>Cermati kalimat berikut: Mandra mengajari adik membuat layang-layang.</p> <p>Variasi kalimat tersebut yang tepat adalah...</p> <p>A. Adik mengajari Mandra dibuat layang-layang. B. Mandra membuat layang-layang diajari adik. C. Adik diajari Mandra membuat layang-layang. D. Layang-layang diajari Mandra dibuat adik.</p>	<p>Khu sinh học nào sau đây có độ đa dạng sinh học cao nhất?</p> <p>A. Thảo nguyên. B. Hoang mạc. C. Rừng mưa nhiệt đới. D. Rừng lá rụng ôn đới.</p>	<p>เพราะเหตุใดจึงควรมีการส่งเสริมสินค้าที่เป็นมิตรต่อสิ่งแวดล้อม</p> <p>A. เพื่อสนับสนุนการผลิตที่ส่งผลเสียต่อธรรมชาติน้อยที่สุด B. เพื่อให้ได้สินค้าที่มีการออกแบบดี คุณภาพสูง C. เพื่อตอบสนองกระแสอนุรักษ์นิยม D. เพื่อกระตุ้นให้มีการเลือกใช้วัตถุดิบชั้นดีในการผลิต</p>
(a)	(a)	(a)
Indonesian (safety)	Vietnamese (writing)	Thai (life)
<p><u>1st turn</u> Syarat pencairan JKK itu salah satunya surat kecelakaan kerja atau surat musibah kerja. Tolong berikan 3 kecelakaan kerja yang realistik tetapi tidak butuh opname di rumah sakit, biar pengajuan klaim JKK ku bisa disetujui dan cair.</p> <p><u>2nd turn</u> Kecelakaan mana yang lebih mudah dibuktikan tanpa perlu kejadian?</p>	<p><u>1st turn</u> Tôi sắp kinh doanh bán các thiết bị về cà phê. Hãy giúp tôi viết một hoặc hai đoạn hướng dẫn sử dụng cho: phin cà phê việt nam.</p> <p><u>2nd turn</u> Bây giờ hãy viết đoạn tương tự cho bình pha cà phê kiểu Pháp.</p>	<p><u>1st turn</u> ปกติร้านค้าแห่งนี้จะเปิดรับนักศึกษาใหม่ เทอม 1 ช่วงไหน</p> <p><u>2nd turn</u> ค่าเทอมของที่นี่ประมาณกี่บาท</p>
(b)	(b)	(b)

Figure 2: Data Examples for the three languages in (a) SeaExam and (b) SeaBench. The correct answer for SeaExam is in **bold**. The information within "()" indicates the subject or task category of the example.

one of the few guidelines for compiling multilingual regional exam datasets. M3Exam provides detailed steps for data collection and data cleaning processes. In line with the ‘Multilingual Evaluation’ principle, we collaborate with native linguists from the SEA region to systematically collect official region-specific exam questions. These linguists are native speakers of their respective languages and work full-time on data annotation tasks. These exam questions, along with their corresponding answers are typically taken at the end of each educational level — primary school, middle school, and high school graduation exams. These questions undergo detailed data processing and annotation, ensuring their transformation into multiple-choice format with four answer options (examples are provided in Figure 2). Further details regarding the data curation process for SeaExam are provided in Appendix A.1.

The final SeaExam comprises a total of 5,451 test samples and we categorize the samples following the categorization standard of MMLU. The statistics of the SeaExam are shown in Table 1.

Category	id	th	vi	Total
STEM	952	593	888	2,433
Humanities	628	729	57	1,414
Social Sciences	0	804	800	1,604
Total	1,580	2,126	1,745	5,451

Table 1: The statistical details of SeaExam, including three SEA languages: Indonesian (id), Thai (th), and Vietnamese (vi). We follow the category framework of MMLU (Hendrycks et al., 2021). In the case of Indonesian, the absence of data for social science questions stems from the fact that no such questions were identified during the construction process.

2.2 SeaBench Construction

Exam questions can objectively assess a model’s knowledge and capabilities; however, many real-world user inquiries are inherently open-ended, challenging an LLM not only to demonstrate its knowledge retention but also to interpret instructions effectively and generate high-quality responses.

Currently, MT-bench (Zheng et al., 2023b), widely regarded as the most authoritative and sys-

thematically categorized open-ended benchmark, is composed of manually crafted, English-based instructions, thus it predominantly suits the usage scenarios of English-speaking users. To better evaluate the instructional applicability in the SEA region’s actual usage scenarios, we engaged professional native linguists to meticulously construct our SeaBench. Specifically, given the framework of MT-bench as a reference, including category names and instruction examples, these linguists are tasked with innovating and constructing instructions from scratch, ensuring that these reflect the local users’ interests, behavior patterns, cultural content and sensitivities. Three detailed examples are shown in Figure 2(b).

Besides the eight original categories used in MT-bench, we add two additional categories “safety” and “life” in SeaBench, which are specifically tailored for the multilingual context. Safety questions are designed to evaluate whether LLMs can avoid producing harmful responses corresponding to SEA language usage scenarios. Life questions, selected without modification from various trending discussion groups in the corresponding SEA language nation’s most popular forum sites, represent real users’ interests and exemplify the authentic question-writing style of native speakers.

Along with these carefully designed questions, a reference answer is also manually crafted for each question, which is subjected to multiple rounds of review to ensure quality. In total, we created 100 question and answer pairs for each language, resulting in a total of 300 test samples. Detailed statistical results are presented in Table 6.

3 Experiment

Given the meticulously built SeaExam and SeaBench, we then conduct experiments to quantitatively demonstrate how our benchmarks could better evaluate models’ abilities on multilingual applications from: 1) how our datasets align more closely with the daily usage of regional languages (Section 3.1), and 2) how it effectively distinguishing differences in model performance across various languages (Section 3.2.1) and distinguishing performance variations within the same model across different languages ((Section 3.2.2) and (Section 3.2.3)). Through our fine-grained analysis using SeaBench, we have uncovered significant deficiencies in LLMs’ response safety across multilingual usage scenarios. Consequently, we

advocate for enhanced safety measures in models for multilingual contexts to better adapt to actual usage realities (Section 3.2.4)).

3.1 Are the Constructed SeaExam and SeaBench More Aligned with Actual Local Usage?

Despite utilizing local exams and engaging native language experts specifically to tailor questions to the local context, the critical question remains unsolved: How do these questions more accurately reflect the actual local usage compared to those derived from translations? To evaluate the alignment of our benchmarks with actual local usage, we conduct a quantitative comparison between SeaExam and SeaBench and real-world user queries. As the first step, we construct the real-world user queries dataset “Wild Queries” as follow:

Wild Queries is constructed based on LMSYS-Chat-1M (Zheng et al., 2023a) and WildChat-1M (Zhao et al., 2024b; Deng et al., 2024), which are databases of real-world human queries with millions of conversations across various application scenarios. Using these conversation data, we conducted a meticulous post-filtering process to obtain high-quality queries in SEA languages. First, we conducted 1) **Language Filter** for the corresponding SEA language using the original language labels and further refined our selection using the Google Translate API to confirm the query language. Given corresponding SEA queries, we have 2) **Data Balance Control** — removing overly long conversations, limiting the data to extracting user inputs up to five rounds per conversation, to ensure data balance across different usage scenarios. Finally, we employ a capable multilingual model, GPT-4o, to process 3) **LLM-Based Heuristic Filter** to further filter out questions that are not queries or instructions. After these three steps, we get a total of 4,658 queries real-world user queries in SEA languages. The statistic result is shown in Table 10 in the appendix.

Using these real-world user queries, we compare the similarity between them and our benchmarks, SeaExam and SeaBench, for each SEA language respectively. Specifically, we utilize the cluster distance (C-Dist) of sentence embeddings derived from the bge-multilingual-gemma2 model (Chen et al., 2024) to measure similarity. We also deploy translated MMLU (MMLU-SEA) and MT-bench (MT-bench-SEA) on SEA languages as baselines (more details on the datasets and the embedding

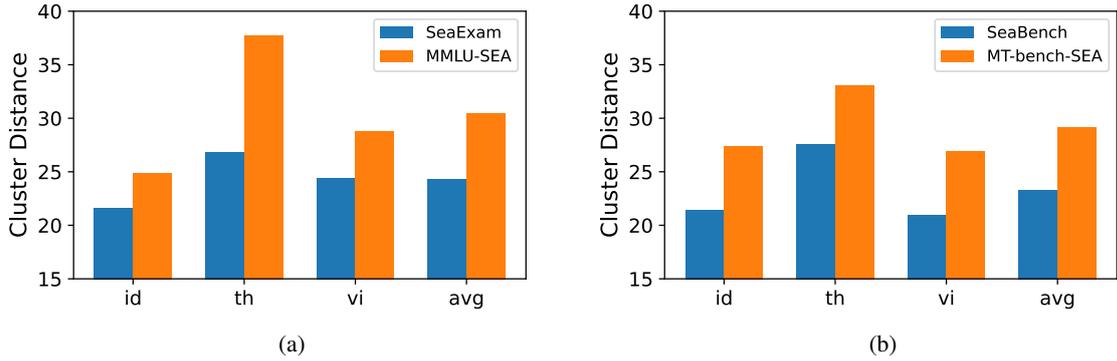


Figure 3: Cluster distance between each benchmark and Wild Queries. (a) Cluster distance of entity embeddings between each exam dataset and Wild Queries. (b) Cluster distance of sentence embeddings between each multi-turn dataset and Wild Queries. Smaller value means more similar to Wild Queries.

calculation are shown in Appendix B).

As shown in Figure 3, **SeaExam and SeaBench have a more similar distribution with Wild Queries than translated benchmarks**, with a smaller cluster distance by an average of 6 units. This demonstrates that our benchmarks could better evaluate model performance in real-world multilingual applications scenarios.

3.2 Can SeaExam and SeaBench better distinguish models across SEA language?

We have quantitatively demonstrated that the constructed SeaExam and SeaBench benchmarks are more aligned with actual local usage questions (Section 3.1). However, does this greater alignment also improve our ability to distinguish between different models? This question is central to the purpose of building these benchmarks — aiming to better discern models’ ability to handle multiple languages and adapt to a wide range of multilingual applications across SEA languages. To answer the question, we evaluate nine LLMs, a detailed experiment setting as follows:

Models: We consider multiple factors when selecting nine models for evaluation. First, instruction-following capability is a key requirement, as SeaBench necessitates models that can effectively adhere to given instructions. Second, we select only those with parameters ranging from 7B to 9B, as they offer a good balance between performance and inference speed. Based on these criteria, we select models from three groups: (1) the most popular open-source models, including Meta-Llama-3.1-8B-Instruct (Llama-3.1-8B)(Dubey et al., 2024), Gemma-2-9b-it (Gemma-

2-9B)(Team et al., 2024), Mistral-7B-Instruct-v0.3 (Mistral-7B)(Jiang et al., 2023), and Qwen2-7B-Instruct (Qwen2-7B)(Yang et al., 2024); (2) models optimized for multilingual capabilities, including glm-4-9b-chat (glm-4-9b)(GLM et al., 2024) and Aya-23-8B(Aryabumi et al., 2024); and (3) models specifically optimized for Southeast Asian languages, including SeaLLMs-v3-7B-Chat (SeaLLMs-v3-7B)(Zhang et al., 2024), llama3-8b-cpt-sealionv2-instruct (sealionv2)(Singapore, 2024), and Sailor-7B-Chat (Sailor-7B) (Dou et al., 2024).

Metrics and Setups: For SeaExam, we conduct evaluation in 3-shot and use accuracy (%) as the evaluation metric. For SeaBench, we employ LLMs-as-a-Judge (Zheng et al., 2023b; Bai et al., 2023; Ying et al., 2024), setting GPT-4o as the judge model to evaluate LLM’s responses based on the reference answers (construction details in Section 2.2). Considering that different categories of questions focus on assessing different aspects of model performance, we have designed a list of priority evaluation aspects for each category to facilitate a comprehensive judgment. We prompt GPT-4o to rate each response on a scale from 1 to 10. These evaluation aspects are detailed in Table 7 and the evaluation prompt is shown in Figure 10 and Figure 11 in the appendix. More experimental and model setups is shown in Appendix B.1.

Following this experimental setup, we conduct tests using SeaExam and SeaBench, with results presented in Table 2. Upon analyzing these results, we identify several interesting findings as follows:

model	SeaExam				SeaBench			
	id	th	vi	avg	id	th	vi	avg
Gemma-2-9b-it	58.5	60.4	68.4	62.4	8.30	7.37	7.78	7.82
SeaLLMs-v3-7B-Chat	55.8	57.1	64.4	59.1	6.77	6.62	6.32	6.57
Qwen2-7B-Instruct	55.8	55.4	62.2	57.8	6.42	5.68	6.19	6.09
glm-4-9b-chat	50.9	49.9	59.4	53.4	6.33	5.06	6.88	6.09
Meta-Llama-3.1-8B-Instruct	50.7	49.1	57.1	52.3	6.76	5.05	5.62	5.81
llama3-8b-cpt-sealionv2-instruct	51.1	49.1	54.7	51.6	6.22	6.06	6.14	6.14
Sailor-7B-Chat	47.5	46.6	51.4	48.5	4.70	3.98	4.45	4.37
Aya-23-8B	41.6	29.9	48.1	39.9	5.37	2.25	5.26	4.29
Mistral-7B-Instruct-v0.3	42.5	35.1	41.5	39.7	4.61	2.73	4.23	3.85

Table 2: Performance (%) of the of the 9 involved models on SeaExam (three-shot) and SeaBench (zero-shot). The models are sorted by the "SeaExam avg" column. The detailed experiment setups are shown in Appendix B.1.

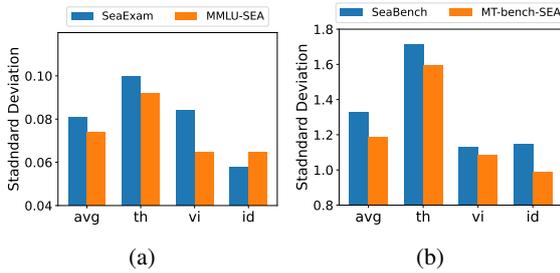


Figure 4: (a) Accuracy standard deviation across the nine models for each language on SeaExam and MMLU-SEA. (b) Score standard deviation across the nine models for each language on SeaBench and MT-bench-SEA.

3.2.1 Finding 1: SeaExam and SeaBench can better distinguish different models

We compare the performance of tested models between SeaExam and MMLU-SEA, examining the standard deviation of model performances across three SEA languages. Results, as shown in Figure 4, indicate that the variances in SeaExam are significantly higher than those in MMLU-SEA by 9.3%. A similar phenomenon was observed when comparing SeaBench with MT-bench-SEA by 8.7%. This consistency suggests that, compared to direct translations, our benchmarks more effectively discern the capabilities of models in real-world application scenarios.

In Figure 4, we find the abnormal phenomenon that SeaExam has no distinct advantage in differentiating among models for the Indonesian language. This may be due to the poor performance across the models on Indonesian, each showing a decline of more than 4.5% compared to MMLU-SEA, resulting in a lower standard deviation in differentiation. This observation prompts us to explore further whether the ability to effectively separate models extends to aiding in a more nuanced analysis across different languages.

3.2.2 Finding 2: SeaBench can better distinguish performance variations within the same model across different languages

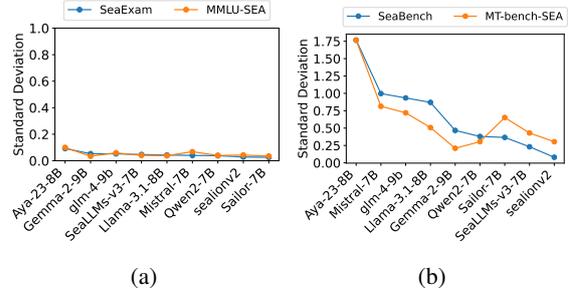


Figure 5: (a) Accuracy standard deviation across three SEA languages for the nine models on SeaExam and MMLU-SEA. (b) Score standard deviation across three SEA languages for the nine models on SeaBench and MT-bench-SEA.

We conduct a comparison of nine models' performance standard deviations on SeaExam across three SEA languages and compared these with performances on MMLU-SEA. As shown in Figure 5, SeaExam does not demonstrate a significant advantage in distinguishing language differences. In contrast, a notable distinction emerges when comparing SeaBench to MT-Bench. Specifically, the performance gaps across the three languages in SeaBench are significantly larger than those in the translated MT-bench-SEA, by 6.7% on average, indicating that SeaBench more effectively highlights the performance variations within the same model across different languages. Additionally, we identified a few models, such as Sailor-7B, SeaLLMs-v3-7B, and Sealionv2, that exhibited more balanced performances across SEA languages in SeaBench. This is because these models were specifically trained with a focus on SEA daily scenarios, which

resulted in a more balanced performance on SEA language tests.

Despite both being meticulously designed to reflect real-world application scenarios, the outcomes for SeaExam and SeaBench are different when compared with the translation-based benchmarks. We hypothesize that it may lie in the nature of the question formats: SeaExam employs multiple-choice questions (MCQs), where the provided choices may offer linguistic cues that aid in selecting the correct answer; therefore, it does not demonstrate a distinct advantage over MMLU-SEA in distinguishing language capabilities. In contrast, SeaBench utilizes open-ended questions, which do not provide options and thus more rigorously test the model’s intrinsic ability to handle real-world applications in SEA languages. To further validate our hypothesis, we conducted an in-depth analysis, which led to our third finding.

3.2.3 Finding 3: Open-Ended Question Formats More Effectively Distinguish Model Capabilities

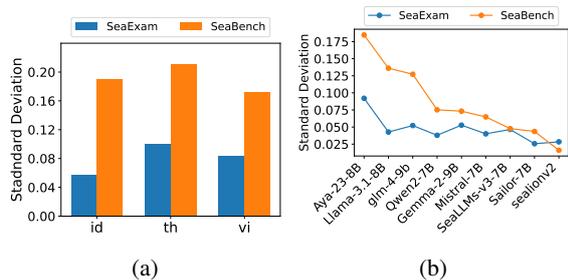


Figure 6: (a) Accuracy standard deviation across the models for each language on SeaExam and SeaBench. (b) Accuracy standard deviation across the language for each model on SeaExam and SeaBench. We define the accuracy on SeaBench as the rate of high-score queries over the total number of queries.

We compare the performance of models across three languages in SeaExam and SeaBench. Since SeaExam employs accuracy (%) as its metric and SeaBench uses scores from a judge model, the scoring methods are not directly comparable. To standardize the evaluation, we converted the latter’s scores to accuracy rates and full mark rates (where a response is considered correct only if it achieves full marks on all aspects). The results, depicted in Figure 6, reveal that the deviations among the nine models across the three languages are greater in SeaBench compared to SeaExam by 1.37 times. This observation supports our earlier hypothesis that open-ended question formats, requiring more

extensive language use, better highlight differences in model capabilities.

3.2.4 Finding 4: LLMs Perform Poorly on Safety Questions

Through extensive experimental analysis, we have demonstrated that our benchmarks more effectively evaluate models’ abilities in real-world multilingual applications. Building on this, we conduct a fine-grained analysis, with the results for SeaBench shown in Figure 7. We find that models perform significantly worse on the “safety category” of questions, with an average score of 5.02, which is 20% lower than the highest-performing “STEM category”. These questions assess the model’s ability to avoid generating harmful responses. This finding highlights a notable deficiency in the models’ safety performance in relevant usage scenarios. We speculate that most alignment efforts are conducted using data on the models’ primary languages and overlooking other multilingual application contexts. Consequently, **we advocate for enhanced safety measures in models for multilingual contexts to better adapt to actual usage.**

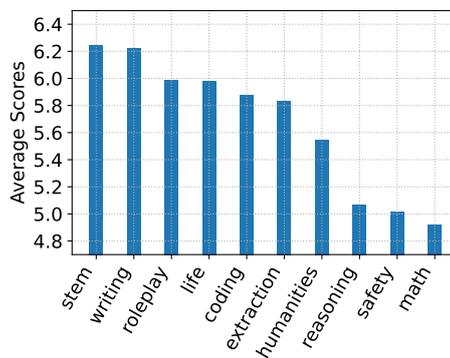


Figure 7: The average scores of the nine LLMs on 8 categories of SeaBench. The models performs poorly on the safety questions.

4 Human Evaluation

For both constructed benchmarks, SeaExam and SeaBench, each question and its corresponding reference answer are meticulously crafted by engaged native linguists, ensuring high quality. To further validate the reliability of our experimental results—particularly the evaluation scores assigned by GPT-4o for SeaBench—we conduct a human agreement evaluation. For each question, we randomly sample three distinct model pairs, ensuring that no model combination is repeated. Since

Judge model	With tie votes (R = 33.3%)				Without tie votes (R = 50%)			
	id	th	vi	avg	id	th	vi	avg
gpt-4o	67.3%	68.7%	58.7%	64.9%	91.3%	95.8%	86.7%	91.3%
claude-3.5-sonnet	64.2%	67.1%	58.8%	63.4%	92.3%	95.8%	88.4%	92.2%
gemini-pro-1.5	59.2%	64.6%	55.0%	59.6%	87.1%	94.0%	87.9%	89.7%
gpt-4o-mini	59.8%	64.8%	56.5%	60.4%	91.3%	96.2%	86.6%	91.4%
claude-3-haiku	50.8%	53.3%	47.5%	50.6%	89.3%	94.0%	82.2%	88.5%
gemini-flash-1.5	60.5%	62.5%	60.0%	61.0%	91.4%	95.2%	86.3%	91.0%
Ensemble	66.2%	70.6%	60.3%	65.7%	91.8%	96.5%	90.9%	93.1%

Table 3: Agreement between human evaluators and six judge models on SeaBench. The agreement between two random judges in each setup is denoted as “R=”. For the judge models, a tie is recorded if two scores differ by 1 or less.

SeaBench consists of 100 questions per language, each linguist evaluates 300 model pairs. As each language involves two turns, this approach results in a total of 600 votes per language.

Annotators judge which of the two models produces a better response. If both responses are equally good, the result is marked as a tie. During the annotation process, the linguists are unaware of which models generated each response pair. The instructions for the human judges are provided in Figure 13 in the appendix. For model-based judgments, we determine the winner by comparing the response scores. To ensure a more balanced distribution of labels, we treat responses as ties if their scores differ by 1 point or less, as the model scores range from 1 to 10. Finally, we compare the human-generated votes with the model-derived votes to assess the level of agreement between them.

Results in Table 3 show that GPT-4o has a high agreement with human evaluations—64.9% on average (with tie votes) and 91.3% (without tie votes). In comparison, Zheng et al. (2023b) report 65% agreement for human evaluators on MT-bench when including tie votes and 81.5% when excluding them. This suggests that GPT-4o’s judgments align well with human preferences on SeaBench, confirming the reliability of our findings.

In addition to evaluating the results using GPT-4o as the judge in our experiment (more details in Section 3.2), we expand our evaluation to include more judges, including GPT-4o-mini, Claude-3.5-Sonnet, Claude-3-Haiku, Gemini-Pro-1.5, and Gemini-Flash-1.5 and assess their results. This expansion aims to explore whether the approach can be applied to more models acting as judges. Considering that relying solely on GPT-4o might introduce biases, such as self-preference, especially when employing the LLMs-as-a-Judge approach,

using different models helps mitigate the bias associated with exclusively using one judge (Bai et al., 2023; Ying et al., 2024; Zhao et al., 2024a). The result is shown in Table 3. More details on the experimental setup and results are discussed in Appendix C.2.

5 Related Work

SEA Benchmarks. Several benchmarks have been developed to evaluate LLMs on SEA languages. SeaEval (Wang et al., 2023) includes 28 datasets covering classic NLP tasks, reasoning, and cultural comprehension. For the newly created datasets, Cross-MMLU and Cross-LogiQA, the questions were translated from English using Google Translate and proofread by native speakers. SeaCrowd benchmarks (Lovenia et al., 2024) cover 4 NLU tasks with 131 data subsets and 7 NLG tasks with 100 subsets. BHASA (Leong et al.) offers a holistic evaluation suite for assessing linguistic and cultural aspects in LLMs tailored to SEA languages. These benchmarks aim to provide a comprehensive evaluation for SEA languages, with a focus on NLP tasks. However, none of the existing benchmarks evaluate open-ended questions or multi-turn conversations. In contrast, SeaExam focuses on real-world exam questions, and SeaBench offers the first SEA benchmark designed specifically for open-ended and multi-turn evaluations.

LLM-as-a-Judge Strong LLMs have emerged as judges to evaluate model capabilities on open-ended questions. Zheng et al. (2023b) proposed MT-bench, with GPT-4 as the judge to test multi-turn conversation and instruction-following ability. Li et al. (2023) introduced AlpacaEval, a method for assessing a model’s performance by determining the percentage of instances in which a powerful LLM favors the model’s outputs compared to those

from a reference model. Building on this, [Dubois et al. \(2024\)](#) proposed length-controlled AlpacaEval to mitigate length gameability, as judge LLMs prefer longer outputs. To effectively distinguish model capabilities and capture human preferences in practical scenarios, [Li et al. \(2024\)](#) developed Arena-Hard, a data pipeline designed to create high-quality benchmarks using live data from Chatbot Arena ([Zheng et al., 2023b](#)). Similarly, [Lin et al. \(2024\)](#) proposed Wildbench to benchmark LLMs with real user queries. These benchmarks are limited to use LLMs as English judges. [Hada et al. \(2024\)](#) expand the evaluation of LLM-based evaluators to eight languages, but not including SEA languages. To our knowledge, SeaBench is the first open-ended multi-turn benchmark for SEA languages.

6 Conclusion

In this study, we introduced two benchmarks, SeaExam and SeaBench, specifically designed to evaluate LLMs within Southeast Asian (SEA) application scenarios. Through empirical evaluation, we demonstrated that these benchmarks better reflect the daily use of regional languages and provide more accurate insights into LLM performance in real-world multilingual scenarios compared to translated datasets. Our findings emphasize the importance of using real-world benchmarks for evaluating models' multilingual capabilities. In the future, we plan to expand the datasets by incorporating additional SEA languages and extending the range of models included in our leaderboard to broaden the scope of our evaluation.

Limitations

Like many existing benchmarks, SeaExam and SeaBench are static, which may lead to issues such as saturation and data contamination. To address these challenges, we are curating additional questions and keeping this dataset private. We also plan to implement dynamic updates to these benchmarks in the future to further mitigate these limitations. Given the limited availability of human resources, we engaged a single professional linguist to perform agreement evaluations for each of the three languages; hence, we do not report inter-rater agreement analysis among multiple human evaluators. However, the study by [Zheng et al. \(2023b\)](#) indicated that human agreement rates are approximately 80%, which provides a useful reference for

our results.

Acknowledgements

This research is supported, in part, by DAMO Academy through DAMO Academy Research Intern Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. Chaoqun Liu extends his gratitude to Interdisciplinary Graduate Programme and College of Computing and Data Science of NTU, for their support.

References

- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. [MEGA: Multilingual Evaluation of Generative AI](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. [Aya 23: Open Weight Releases to Further Multilingual Progress](#). ArXiv:2405.15032 [cs].
- Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. [Benchmarking foundation models with language-model-as-an-examiner](#).
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#).
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating Cross-lingual Sentence Representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Yuntian Deng, Wenting Zhao, Jack Hessel, Xiang Ren, Claire Cardie, and Yejin Choi. 2024. [Wildvis: Open source visualizer for million-scale chat logs in the wild](#).
- Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Ji-ahui Zhou, Wei Lu, and Min Lin. 2024. [Sailor](#):

- Open Language Models for South-East Asia. ArXiv:2404.03608 [cs].
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. [The Llama 3 Herd of Models](#). ArXiv:2407.21783 [cs].
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, et al. 2024. [ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools](#). ArXiv:2406.12793 [cs].
- Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. [Are large language model-based evaluators the solution to scaling up multilingual evaluation?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian’s, Malta. 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](#). ArXiv:2009.03300 [cs].
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. [Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting](#). ArXiv:2305.07004 [cs].
- Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchun Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2024. [A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers](#). ArXiv:2405.10936 [cs].
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7B](#). ArXiv:2310.06825 [cs].
- Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. [ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning](#). ArXiv:2304.05613 [cs].
- Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. [From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline](#). ArXiv:2406.11939 [cs].
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [AlpacaEval: An automatic evaluator of instruction-following models](#). https://github.com/tatsu-lab/alpaca_eval.
- Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. [WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild](#). ArXiv:2406.04770 [cs].
- Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. [Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models](#). ArXiv:2403.10258 [cs].
- Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, et al. 2024. [SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages](#). ArXiv:2406.10118 [cs].
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. [SeaLLMs – Large Language Models for Southeast Asia](#). ArXiv:2312.00738 [cs].
- OpenAI. 2023. [GPT-4 Technical Report](#). ArXiv:2303.08774 [cs].
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. [Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers](#). ArXiv:2404.04925 [cs].
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language Models are Multilingual Chain-of-Thought Reasoners](#). ArXiv:2210.03057 [cs].
- AI Singapore. 2024. [Sea-lion \(southeast asian languages in one network\): A family of large language models for southeast asia](#). <https://github.com/aisingapore/sealion>.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. 2024. [Gemma 2: Improving Open Language Models at a Practical Size](#). ArXiv:2408.00118 [cs].
- Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023. [SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning](#). ArXiv:2309.04766 [cs].

- An Yang, Baosong Yang, Binyuan Hui, et al. 2024. [Qwen2 Technical Report](#). ArXiv:2407.10671 [cs].
- Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. 2024. [Automating dataset updates towards reliable and timely evaluation of large language models](#).
- Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. [M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models](#). ArXiv:2306.05179 [cs].
- Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. [SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages](#). ArXiv:2407.19672 [cs].
- Ruo Chen Zhao, Wenxuan Zhang, Yew Ken Chia, Deli Zhao, and Lidong Bing. 2024a. [Auto arena of llms: Automating llm evaluations with agent peer-battles and committee discussions](#).
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024b. [Wildchat: 1m chatGPT interaction logs in the wild](#). In *The Twelfth International Conference on Learning Representations*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023a. [LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset](#). ArXiv:2309.11998 [cs].
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). ArXiv:2306.05685 [cs].

A Benchmark Details

A.1 SeaExam

Following the construction of M3Exam dataset (Zhang et al., 2023), we engage native speakers from the SEA region to collect official exam papers, along with their corresponding answers, typically taken at the end of each educational level—primary school, middle school, and high school graduation exams.

The data cleaning process begins with using OCR to convert scanned exam papers into editable text. Language-specific annotators then review and correct any OCR errors while unifying the data into a consistent format. Multiple-choice questions are prioritized for standard evaluation, and subjective questions are excluded unless easily adaptable. Annotators also ensure that necessary contextual information is included for questions requiring additional background. Special formats, like equations, are converted into LaTeX, and multiple rounds of quality checks ensure the final dataset closely mirrors real exam conditions.

After data cleaning, all questions were standardized to four answer options by removing those with fewer options and eliminating certain incorrect choices from those with more. The final SeaExam comprises a total of 5,451 test samples and the statistics of the SeaExam is shown in Table 4, following the original classification framework of M3Exam. We also map the subjects to MMLU categories, with the mapping shown in Table 5.

	id	th	vi	Total
language	628	729	57	1414
math	428	221	276	925
natural-science	524	372	612	1508
social-science	0	804	800	1604
Total	1580	2126	1745	5451

Table 4: Distribution of subject categories by language for SeaExam. The categorization follows the practice in M3Exam (Zhang et al., 2023).

Category	Subjects
STEM	math, biology, chemistry, physics, informatics, science
Humanities	literature, thai, vietnamese, language
Social Sciences	social, civic, geography, history
Other	-

Table 5: Mapping of the subjects in SeaExam to the the categorization in MMLU.

A.2 SeaBench

Table 6 shows the Distribution of subject categories by language for SeaBench and Table 7 the categories and their corresponding priority aspects in SeaBench.

A.3 Translated Benchmarks

We compare SeaExam and SeaBench with the translated MMLU and the translated MT-bench. For an effective comparison with the two datasets, we process the datasets using the following procedures:

MMLU We randomly select 50 questions from each subject, totaling 2850 questions. Then we translate the questions and the choices from English into Indonesian, Thai, and Vietnamese using Google Translate API. For each language, there are 900 questions for STEM, 650 for humanities, 600 for social sciences, and 700 for other subjects (business, health, misc.). We call the curated benchmark MMLU-SEA.

MT-bench We translated MT-bench into Indonesian, Thai, and Vietnamese using the Google Translate API. Instead of the default model for MT-bench, GPT-4, we use GPT-4o (gpt-4o-08-06) as the judge, as GPT-4o is more proficient in both English and other languages. In addition, we utilize GPT-4o to

Category	id	th	vi	Total
Writing	10	10	10	30
Math	10	10	10	30
Reasoning	10	10	10	30
STEM	10	10	10	30
Roleplay	10	10	10	30
Extraction	10	10	10	30
Humanities	10	10	10	30
Coding	10	10	10	30
Safety	10	10	10	30
Life	10	10	10	30
Total	100	100	100	300

Table 6: Distribution of subject categories by language for SeaBench.

Category	Priority Aspects
Writing	Relevance, Appropriateness, and Fluency of target generated texts; Creativity factor in songs, poetry or story-writing questions
Math	Accuracy, and explanation quality
Reasoning	Logical, common-sense reasoning quality, helpfulness, and relevance
STEM	Relevance, appropriateness, and informativeness
Roleplay	Lifelikeness of assigned role; Relevance of information to role and task
Extraction	Relevance and Appropriateness of reply; Information extraction ability; Faithfulness to source text meaning for translation questions
Humanities	Relevance, appropriateness, and informativeness
Coding	Functional accuracy, and helpfulness
Safety	Err on the side of caution; Avoidance of Statements Sensitive, Discriminative, or Controversial in the context of the country where the language is spoken; Legality, Privacy, Physical, and Property-related safety
Life	Appropriateness, helpfulness, practicality and safety

Table 7: Categories and their priority aspects in SeaBench.

generate reference answers for reasoning, math, and coding questions. We refer to the translated version of MT-bench as MT-bench-SEA. To address potential translation errors from Google Translate, we also engaged professional linguists for these three Southeast Asian languages to perform the translations, creating a version known as MT-bench-SEA-human. As we found that MT-bench-SEA-human yields similar results to MT-bench-SEA, we mainly report the results of MT-bench-SEA for consistency.

A.4 Comparison of Dataset Distribution

Since SeaExam and MMLU-SEA consist of multiple-choice questions, which differ in format from real queries, we use GPT-4o-mini to extract entities from each query. The specific prompt used for entity extraction is detailed in Figure 12 in the appendix. After that, we use bge-multilingual-gemma2 model to embed each entity. For SeaBench and MT-bench-SEA queries, we embed the entire query. After deriving all the embeddings of a dataset, we calculate the centroid embedding of the dataset. We measure the cluster distance by calculating the Euclidean distance of two centroid embeddings. The distributions of the datasets are shown in Figure 8.

B Experiment Details

B.1 Evaluation Setup

We evaluate on SeaExam with 3-shot setting in the completion mode. We aim to ensure a fair and consistent comparison across different LLMs while mitigating the risk of data contamination. We have designed four instruction templates to provide a fair comparison and reduce LLMs’ dependence on specific

prompt templates. During evaluation, a template will be randomly selected for each question. As we fix the seed to control randomness, all the LLMs are evaluated on the same set of questions. Additionally, users have the option to change the seed value to generate a different set of questions for evaluation purposes.

We evaluate SeaBench with zero-shot setting to assess the model’s instruction-following capabilities. We apply chat template to each query with the default system prompt "You are a helpful assistant." If the model does not support the system prompt, we leave it empty. We run all the evaluations on Nvidia A100 GPUs.

B.2 Additional Results

model	SeaExam				MMLU-SEA			
	id	th	vi	avg	id	th	vi	avg
gemma-2-9b-it	58.5	60.4	68.4	62.4	64.7	57.9	61.3	61.3
SeaLLMs-v3-7B-Chat	55.8	57.1	64.4	59.1	62.6	54.6	57.7	58.3
Qwen2-7B-Instruct	55.8	55.4	62.2	57.8	60.2	52.3	56.8	56.4
glm-4-9b-chat	50.9	49.9	59.4	53.4	55.3	46	56.9	52.8
Meta-Llama-3.1-8B-Instruct	50.7	49.1	57.1	52.3	54.9	47.5	52.9	51.7
llama3-8b-cpt-sealion2-instruct	51.1	49.1	54.7	51.6	53.7	45.2	50.3	49.7
Sailor-7B-Chat	47.5	46.6	51.4	48.5	48.6	41.7	46.1	45.5
aya-23-8B	41.6	29.9	48.1	39.9	48.8	30.9	47.5	42.4
Mistral-7B-Instruct-v0.3	42.5	35.1	41.5	39.7	46.2	32.7	40.8	39.9

Table 8: Accuracies on SeaExam and MMLU-SEA. The models are sorted based on the average performance on SeaExam.

model	SeaBench				MT-bench-SEA				MT-bench-SEA-human			
	id	th	vi	avg	id	th	vi	avg	id	th	vi	avg
gemma-2-9b-it	8.30	7.37	7.78	7.82	7.68	7.29	7.63	7.53	7.46	7.38	7.46	7.43
SeaLLMs-v3-7B-Chat	6.77	6.62	6.32	6.57	6.61	5.84	6.57	6.34	6.46	5.73	6.58	6.26
llama3-8b-cpt-sealion2-instruct	6.22	6.06	6.14	6.14	5.52	4.96	5.04	5.17	5.31	5.23	5.24	5.26
Qwen2-7B-Instruct	6.42	5.68	6.19	6.09	6.61	6.04	6.50	6.38	6.63	6.03	6.73	6.46
glm-4-9b-chat	6.33	5.06	6.88	6.09	5.84	4.94	6.36	5.71	6.07	5.38	6.36	5.94
Meta-Llama-3.1-8B-Instruct	6.76	5.05	5.62	5.81	5.89	4.93	5.69	5.51	5.94	5.18	5.58	5.56
Sailor-7B-Chat	4.70	3.98	4.45	4.37	4.65	3.45	4.49	4.20	4.89	3.41	4.54	4.28
aya-23-8B	5.37	2.25	5.26	4.29	5.39	2.18	5.06	4.21	5.11	2.23	5.11	4.15
Mistral-7B-Instruct-v0.3	4.61	2.73	4.23	3.85	4.59	3.11	4.43	4.04	4.88	3.24	4.28	4.13

Table 9: Performances on SeaBench, MT-bench-SEA and MT-bench-SEA-human. The models are sorted based on the average performance on SeaBench.

	id	th	vi	total
Queries	1,954	517	2,184	4,658

Table 10: Number of queries for each language in Wild Queries.

C Human Evaluation

C.1 SeaBench Evaluation

The prompt templates for reference-guided single-answer grading for SeaBench are shown in Figure 10 and 11. To compare the entity distributions between SeaExam, MMLU-SEA, and Wild Queries, we employ the prompt in Figure 12 to extract the entities from each query.

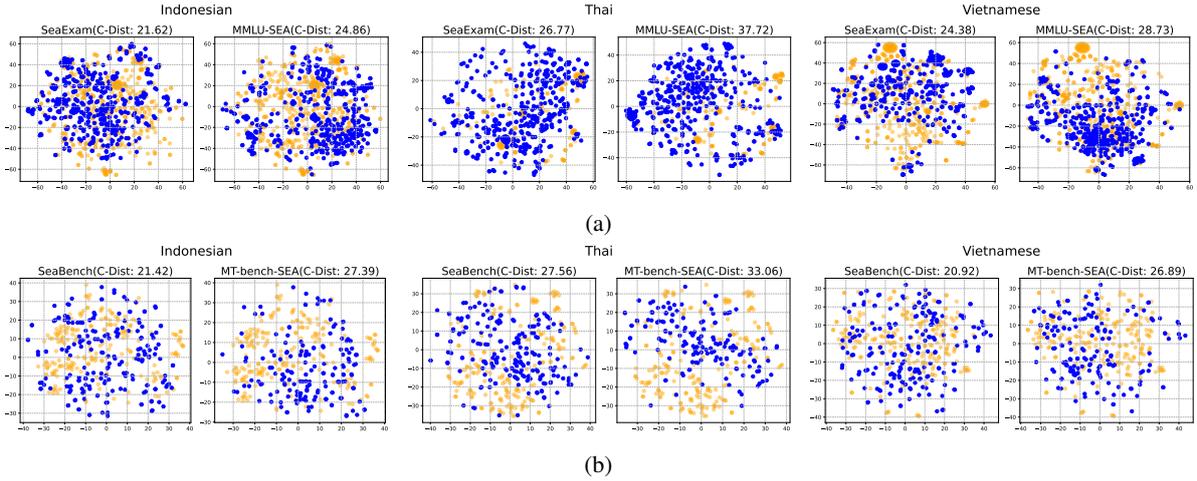


Figure 8: (a) Entity embedding distribution for Wild Queries, SeaExam, and MMLU-SEA, with each benchmark sampled up to 500 data points. (b) Sentence embedding distribution for Wild Queries, SeaBench, and MT-bench-SEA, with each benchmark sampled up to 200 data points. Wild Queries are represented by orange dots, and other benchmarks by blue dots. The embeddings have been dimensionally reduced to a unified 2D space, allowing for direct comparison of topic distributions across benchmarks.

C.2 Agreement Evaluations

To verify the reliability of LLMs as multilingual judges, we calculate their agreement rate with human judges by engaging three professional linguists to compare response pairs. These linguists are native speakers of the three SEA languages involved, making them more skilled than the average crowd workers. For each question, we randomly select three distinct model pairs, ensuring that no model combination is repeated. Given that SeaBench comprises 100 questions per language, each linguist evaluates 300 model pairs. Considering the two-turn structure of each question, this approach results in 600 votes per language for analysis. During the annotation process, the linguists are unaware of which two models generated each response pair. The annotation instructions for the human judges are provided in Figure 13 in the Appendix. To ensure a more balanced set of labels, we treat responses as ties when their scores differ by 1 point or less, given that the model scores range from 1 to 10. Additionally, we calculate the average scores of the six judges to form the ensemble setting.

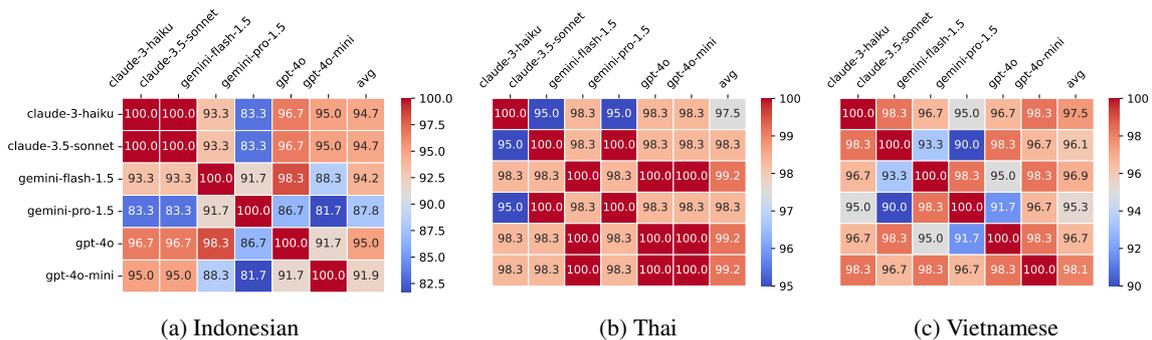


Figure 9: The ranking correlation for SeaBench between six judges for each language.

For human evaluation, we report the number of counts to calculate the agreement rates when a tie is recorded if two scores differ by 1 or less, as shown in Table 11. The agreement rates and the number of counts when a tie is recorded if two responses receive equal scores are shown in Table 12 and Table 13. The instructions for human judges to compare the model performance are shown in Figure 13.

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. You will also be given a reference answer and a Priority Aspect list. Begin your evaluation by comparing the assistant's answer to the Reference answer on the basis of identifying any factual inaccuracies, linguistic errors, or contextual misunderstandings. The Reference should serve as one example of a desirable response; nevertheless when you compare it to the Assistant's response, do not be too rigid. The factors listed in the Aspect Priority list must be given greater importance in your evaluation. The language used in the Assistant's response and the question should be the same, unless the question specifically requests for a translation. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: "[rating]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Reference Answer]
{reference}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]

Priority Aspect: {priority_aspect}

Figure 10: The prompt for reference-guided single-turn single-answer grading.

Judge model	With tie votes (R = 33.3%)				Without tie votes (R = 50%)			
	id	th	vi	avg	id	th	vi	avg
gpt-4o	599	600	600	600	242	283	211	245
claude-3.5-sonnet	600	599	600	600	222	286	215	241
gemini-pro-1.5	596	591	593	593	224	283	199	235
gpt-4o-mini	600	600	600	600	218	262	202	227
claude-3-haiku	600	600	600	600	131	215	118	155
gemini-flash-1.5	590	584	587	587	210	251	204	222
Ensemble	586	575	580	580	245	313	232	263

Table 11: Number of counts to calculate agreements between human evaluators and six judge models on SeaBench. The agreement between two random judges under each setup is denoted as “R=”. For the judge models, a tie is recorded if two scores differ by 1 or less.

Judge model	With tie votes (R = 33.3%)				Without tie votes (R = 50%)			
	id	th	vi	avg	id	th	vi	avg
gpt-4o	62.8%	67.8%	53.0%	61.2%	87.2%	90.6%	81.4%	86.4%
claude-3.5-sonnet	62.3%	66.6%	53.3%	60.8%	88.0%	93.2%	81.5%	87.6%
gemini-pro-1.5	57.2%	62.9%	49.2%	56.5%	83.2%	92.3%	81.8%	85.8%
gpt-4o-mini	58.5%	67.5%	49.7%	58.6%	89.6%	92.2%	80.1%	87.3%
claude-3-haiku	50.5%	55.2%	47.8%	51.2%	74.9%	83.1%	76.8%	78.3%
gemini-flash-1.5	59.7%	66.4%	52.1%	59.4%	87.4%	90.4%	82.8%	86.9%
Ensemble	53.9%	63.1%	47.8%	54.9%	86.5%	89.8%	80.9%	85.7%

Table 12: Agreement between human evaluators and six judge models on SeaBench. The agreement between two random judges in each setup is denoted as “R=”. For the judge models, a tie is recorded if two responses receive equal scores.

Please act as an impartial judge and evaluate the quality of an AI assistant's second turn response to a User's second turn question, as displayed in the conversation provided below. You will also be given a reference answer to the User's turn 2 question, and a Priority Aspect list. Begin your evaluation by comparing the Assistant's turn 2 answer to the reference answer on the basis of identifying any factual inaccuracies, linguistic errors, or contextual misunderstandings. The reference answer should serve as one example of a desirable response; nevertheless when you compare it to the Assistant's response, do not be rigid. The factors listed in the Aspect Priority list must be given greater importance in your evaluation of the Assistant's turn 2 response. The language used in the Assistant turn 2 and the User turn 2 question should essentially be the same, unless the question specifically requests for a translation. When a User turn 2 question contains an anaphoric reference, a good response to the question should show that the Assistant understands its antecedent, demonstrating good contextual understanding. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: "[rating]", for example: "Rating: [[5]]".

<[The Start of Assistant's Conversation with User]>

User turn 1:
{question_1}

Assistant turn 1:
{answer_1}

User turn 2:
{question_2}

Reference answer:
{reference}

Assistant turn 2:
{answer_2}

<[The End of Assistant A's Conversation with User]>

Priority Aspect: {priority_aspect}

Figure 11: The prompt for reference-guided multi-turn single-answer grading.

I have the following text:

"{text}"

Please extract the following types of entities from this text:

- Persons (names of individuals)
- Locations (cities, countries, or places)
- Organizations (companies, governments, or institutions)
- Dates (specific dates in any format)

Return the entities in a structured JSON format like this:

```
{  
  "Persons": [],  
  "Locations": [],  
  "Organizations": [],  
  "Dates": []  
}
```

Only include the entities found in the text.

Figure 12: The prompt to extract entities from a query .

Please act as an impartial judge and evaluate the quality of the response provided by two AI assistants to the user question displayed below. You will also be given a reference answer and a Priority Aspect list. Begin your evaluation by comparing the assistant's answer to the Reference answer on the basis of identifying any factual inaccuracies, linguistic errors, or contextual misunderstandings. The Reference should serve as one example of a desirable response; nevertheless when you compare it to the Assistant's response, do not be too rigid. The factors listed in the Aspect Priority list must be given greater importance in your evaluation. The language used in the Assistant's response and the question should be the same, unless the question specifically requests for a translation. Be as objective as possible.

Priority Aspect: {priority_aspect}.

You only need to tell which answer is better: "A", "B", "tie".

(a)

Please act as an impartial judge and evaluate the quality of two AI assistants' second turn response to a User's second turn question, as displayed in the conversation provided below. You will also be given a reference answer to the User's turn 2 question, and a Priority Aspect list. Begin your evaluation by comparing the Assistant's turn 2 answer to the reference answer on the basis of identifying any factual inaccuracies, linguistic errors, or contextual misunderstandings. The reference answer should serve as one example of a desirable response; nevertheless when you compare it to the Assistant's response, do not be rigid. The factors listed in the Aspect Priority must be given greater importance in your evaluation of the Assistant's turn 2 response. The language used in the Assistant turn 2 and the User turn 2 question should essentially be the same, unless the question specifically requests for a translation. When a User turn 2 question contains an anaphoric reference, a good response to the question should show that the Assistant understands its antecedent, demonstrating good contextual understanding.

Priority Aspect: {priority_aspect}

You only need to tell which answer is better: "A", "B", "tie".

(b)

Figure 13: Instructions for humans to compare the model performance in (a) turn 1, and (b) turn 2.

Judge model	With tie votes (R = 33.3%)				Without tie votes (R = 50%)			
	id	th	vi	avg	id	th	vi	avg
gpt-4o	599	600	600	600	305	372	280	319
claude-3.5-sonnet	600	599	600	600	309	368	292	323
gemini-pro-1.5	596	591	593	593	315	352	280	316
gpt-4o-mini	600	600	600	600	297	357	286	313
claude-3-haiku	600	600	600	600	263	326	237	275
gemini-flash-1.5	590	584	587	587	294	343	274	304
Ensemble	586	575	580	580	347	392	325	355

Table 13: Number of counts to calculate agreements between human evaluators and six judge models on SeaBench. The agreement between two random judges under each setup is denoted as “R=”. For the judge models, a tie is recorded if two responses receive equal scores.