SEA-HELM:
Southeast Asian Holistic Evaluation of Language Models

Yosephine Susanto[1,2], Adithya Venkatadri Hulagadri[1,2,*], Jann Railey Montalan[1,2,*], Jian Gang Ngui[1,2,*], Xian Bin Yong[1,2,*], Weiqi Leong[1,2], Hamsawardhini Rengarajan[1,2], Peerat Limkonchotiwat[1,2], Yifan Mai[3], William Chandra Tjhi[1,2]

[1]AI Singapore, [2]National University of Singapore
[3]Center for Research on Foundation Models (CRFM), Stanford University
[*]Core contributors
Correspondence: yosephine@aisingapore.org

Abstract

With the rapid emergence of novel capabilities in Large Language Models (LLMs), the need for rigorous multilingual and multicultural benchmarks that are integrated has become more pronounced. Though existing LLM benchmarks are capable of evaluating specific capabilities of LLMs in English as well as in various mid- to low-resource languages, including those in the Southeast Asian (SEA) region, a comprehensive and culturally representative evaluation suite for the SEA languages has not been developed thus far. Here, we present **SEA-HELM**,[1] a holistic linguistic and cultural LLM evaluation suite that emphasises SEA languages, comprising five core pillars: (1) NLP CLASSICS, (2) LLM-SPECIFICS, (3) SEA LINGUISTICS, (4) SEA CULTURE, (5) SAFETY. SEA-HELM currently supports Filipino, Indonesian, Tamil, Thai, and Vietnamese. We also introduce the **SEA-HELM leaderboard**,[2] which allows users to understand models' multilingual and multicultural performance in a systematic and user-friendly manner. We make the SEA-HELM evaluation code publicly available.[3]

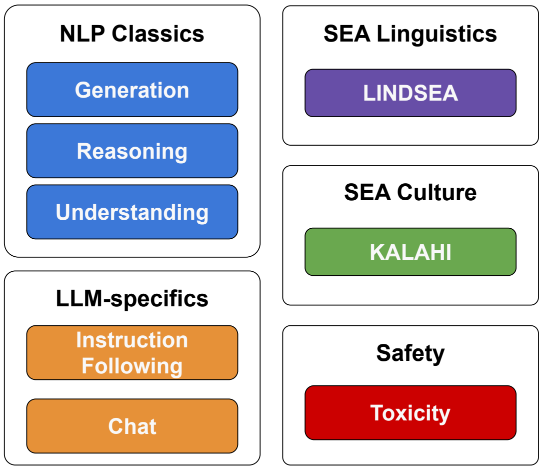


Figure 1: The five evaluation pillars of SEA-HELM that make up our holistic and integrated approach.

1 Introduction

The proliferation of generative approaches to natural language processing (NLP) through Large Language Models (LLMs) has rendered many traditional datasets for NLP evaluation compromised (Haimes et al., 2024), obsolete or saturated (Liu et al., 2024). While essentially trained to predict the next token in a sequence, LLMs have shown significant emergent competencies, including summarisation, question answering, translation, coding, and advanced reasoning (Brown et al., 2020; Yeo et al., 2024). They are also extensively used for new applications, such as chatbots that can hold sustained open-ended conversations (Dam et al., 2024). This advancement has led to a significant disparity between the range of LLM capabilities and the datasets and frameworks to evaluate them rigorously. Traditional approaches to NLP evaluation, which emphasise on alignment with a predefined ground truth reference, are not sufficient in measuring the complex abilities of LLMs (van Schaik and Pugh, 2024; Gema et al., 2025; Grusky, 2023). This gap is further exacerbated in lower-resource languages in **S**outh**E**ast **A**sia (**SEA**), owing to a lack of both training and testing data on the internet (Li et al., 2024).

The SEA region is home to nearly 700 million speakers across more than 1,000 languages (see Appendix H.18). While this region represents almost 10% of the global population and constitutes approximately one-seventh of the world's total languages, most of these languages remain unsupported by major LLMs, such as Mistral (Mistral AI, 2023) and Claude (Anthropic, 2023).

[1]Formerly known as BHASA (Leong et al., 2023).
[2]`https://leaderboard.sea-lion.ai/`
[3]`https://github.com/aisingapore/SEA-HELM`

Findings of the Association for Computational Linguistics: ACL 2025, pages 12308–12336
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

The complications arising from a lack of data as well as uneven digital access and representation contribute to the impeded development of LLMs in those languages in the SEA region. This matter is further aggravated by the fact that some SEA languages are written in non-Latin scripts, which presents a challenge for tokenizers when processing limited data.

Despite the mentioned obstacles, multilingual LLM and benchmark development in SEA strive to close the gap and adapt to the current trends in the field. Some models now explicitly support SEA languages and also claim to provide representation for SEA cultural knowledge (Sailor2 Team, 2024; Zhang et al., 2024b; Bai et al., 2023; Dang et al., 2024). There have also been many benchmarks claiming to measure LLMs' multilingual and multicultural capabilities for the SEA region (Nguyen et al., 2024; Zhang et al., 2024a; Wang et al., 2023; DAMO-NLP-SG, 2024; Singh et al., 2024; CohereForAI, 2024; Lovenia et al., 2024). However, there has yet to be a comprehensive and culturally representative benchmark suite for evaluating LLMs for SEA cultural and linguistic competencies.

We have thus developed **SEA-HELM** (**S**outh**E**ast **A**sian **H**olistic **E**valuation of **L**anguage **M**odels),[4] a systematic, integrated, and continually maintained benchmark suite which aims to measure the SEA language and cultural competencies of LLMs in a targeted and comprehensive manner. SEA-HELM achieves integration by collating localised evaluation datasets and LLM prompts, running tests on models together to enable standardised comparisons, and presenting results aggregated by languages, tasks and models. It is our view that no single metric explains a model's suitability for SEA, and thus SEA-HELM is designed to test a holistic set of competencies, illustrated in Figure 1.

Specifically, SEA-HELM is organised into five evaluation pillars: (1) NLP CLASSICS; (2) LLM-SPECIFICS; (3) SEA LINGUISTICS; (4) SEA CULTURE; and (5) SAFETY, which together encompass an extensive range of tasks for each SEA language to ensure that a wide range of relevant aspects, from linguistic nuances to cultural representations, are considered and accounted for. The five pillars are also meticulously and rigorously crafted to achieve fair, transparent, and authentic multilingual and multicultural evaluations of LLMs in the region. We deliberately incorporate community participation by involving native speakers of the SEA languages at each stage of the dataset planning and construction to ensure linguistic accuracy and cultural authenticity.

We summarise the contributions of **SEA-HELM** here:

- SEA-HELM is a curated suite of SEA datasets which are evaluated together and whose results will be presented on a publicly visible leaderboard.
- SEA-HELM (a) incorporates natively adapted and human-translated benchmark tasks from a range of SEA languages such as Filipino,[5] Indonesian, Tamil,[6] Thai and Vietnamese; (b) created the human-translated SEA-IFEval and SEA-MTBench datasets; and (c) constructed LLM prompt templates in each native language that are consistent across tasks.
- SEA-HELM includes new datasets we developed for granular linguistic diagnostics (LINDSEA) in Indonesian and Tamil. SEA-HELM also includes a cultural evaluation dataset for Filipino, KALAHI, developed in collaboration with community members from the Philippines (Montalan et al., 2024).

We believe that these make SEA-HELM a comprehensive and representative evaluation suite for SEA languages. It can be used as a base for future extensions covering other SEA languages, e.g. Khmer, Lao, Burmese, etc., which we intend to expand on in a future work.

2 Related work

2.1 LLM evaluations

Over the years, AI practitioners have employed either an individual task-based or, more rarely, a holistic approach to assess the performance and capabilities of LLMs. Popular tasks for evaluating LLMs include translation (Hendy et al., 2023), summarisation (Zhang et al., 2023), decision-making (Shen et al., 2023), detecting scalar implicatures (Jeretič et al., 2020; Pandia et al., 2021; Hu et al., 2023; Liu et al., 2023) as

[4] We release SEA-IFEval, SEA-MTBench and LINDSEA datasets under the Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA 4.0) license. This respects the licenses of the source datasets used in this study.

[5] Which also culminated in BATAYAN, the Filipino expansion of SEA-HELM.

[6] Tamil is one of the official languages of Singapore and is also spoken in Malaysia.

well as presuppositions (Jeretič et al., 2020; Parrish et al., 2021). Additionally, linguistic (Warstadt et al., 2020; Xiang et al., 2021; Someya and Oseki, 2023) and cultural representation (Durmus et al., 2023; Atari et al., 2023) are also increasingly recognised as essential criteria for evaluating the efficacy and fairness of language models.

On a holistic approach, Stanford University introduced HELM (Liang et al., 2022) as an initiative aimed at evaluating LLMs across a wide range of tasks, such as linguistic capabilities, reasoning, knowledge, memorisation, disinformation, bias and toxicity. Google introduced BIG-Bench (Srivastava et al., 2023), which is a crowdsourced initiative. Similarly, OpenAI launched OpenAI Evals,[7] a crowdsourced system that invites users to create custom evaluation datasets.

2.2 LLM evaluations for SEA languages

Recently, an increasing amount of attention has been directed towards LLM training and evaluations beyond English. There has been a growing body of work (Sailor2 Team, 2024; Zhang et al., 2024b; Bai et al., 2023; Dang et al., 2024) evaluating the performance of LLMs in a wide range of tasks in SEA languages. Most of them attempted to incorporate a broad spectrum of languages (e.g. Indonesian, Thai, Filipino). In order to achieve such large language coverage, machine translation and synthetic generation were typically used to generate multilingual benchmark datasets.

However, the use of machine translation and synthetically generated benchmarks with little input from the community raises questions on their cultural authenticity and reliability. Automatic translation often misses the cultural nuances inherent in the target language, and can result in translation errors and biases (Singh et al., 2024). This can result in cultural erasure, furthering stereotypical or non-diverse views (Qadri et al., 2025). Thus, there is a need to develop authentic, human-verified multilingual evaluation datasets and metrics. Works such as Singh et al. (2024), Romero et al. (2024), and Koto et al. (2024) address the above point by adopting a participatory framework (Birhane et al., 2022; Smart et al., 2024). The participatory framework is also core to SEA-HELM's design philosophy as it ensures linguistic accuracy and cultural authenticity.

[7] https://github.com/openai/evals/

3 SEA-HELM

To address the lack of holistic multilingual and multicultural evaluations for the SEA region, we designed and developed SEA-HELM, which draws its inspiration from HELM (Liang et al., 2022). This evaluation suite consists of five core pillars: (1) NLP CLASSICS, (2) LLM-SPECIFICS, (3) SEA LINGUISTICS, (4) SEA CULTURE, and (5) SAFETY, and has been recently integrated with HELM. The spread of tasks and languages is detailed in Table 1. SEA-HELM currently supports five SEA languages – Filipino, Indonesian, Tamil, Thai, and Vietnamese, enabling users and AI practitioners to assess the overall performance of LLMs for these languages.

3.1 Core pillars

3.1.1 NLP classics

First, for the Natural Language Understanding (NLU) competency, we include QA (extractive question answering) and sentiment analysis tasks. Second, for the Natural Language Generation (NLG) competency, we include translation (English to native language and native language to English) and abstractive summarisation tasks. Third, for the Natural Language Reasoning (NLR) competency, we include causal reasoning and natural language inference (NLI) tasks.[8]

We selected datasets that comprised of data originally written in the native language as far as possible. Otherwise, existing datasets in English were carefully translated by native speakers.[9] This is important because translated datasets often contain elements of translationese (Gellerstam, 1986), which can differ significantly from natively written text (Baker, 1993; Lembersky et al., 2012; Volansky et al., 2015; Riley et al., 2020).

3.1.2 LLM-specifics

With LLMs enabling unprecedented NLP applications, there is a need to develop automated, dedicated evaluation metrics for these higher-order tasks. SEA-HELM focuses on two specific capabilities - the ability to follow human

[8] It should be noted that reasoning as a competency is broad and often a reasoning task requires many specific types of reasoning skills (Espejel et al., 2023; Huang and Chang, 2023; Qiao et al., 2023; Xu et al., 2023; Yu et al., 2023).

[9] Annotator demographics are not included in this paper for anonymity, demographics can be provided upon request.

Pillar	Competency	Task	Dataset	Language	Metrics	Native	Translation	Our contribution
NLP Classics	NLU	Sentiment	PH Elections Sentiment (Cabasag et al., 2019)	FIL	WA	Y	-	
			NusaX (Winata et al., 2023)	ID	WA	Y	-	
			IndicSentiment (Doddapaneni et al., 2023)	TA	WA	N	Human	
			Wisesight (Bact' et al., 2019)	TH	WA	Y	-	
			UIT-VSFC (Nguyen et al., 2018)	VI	WA	Y	-	
		QA	TyDi QA-GoldP (Clark et al., 2020)	ID	F1	Y	-	
			IndicQA (Doddapaneni et al., 2023)	TA	F1	Y	-	
			XQUAD (Artetxe et al., 2020)	TH, VI	F1	N	Human	
		QA-Multiple Choice	Belebele (Bandarkar et al., 2024)	FIL	F1	N	Human	
		Metaphor	MABL (Kabra et al., 2023)	ID	WA	Y	-	
	NLR	NLI	XNLI (Conneau et al., 2018)	FIL, TH, VI	WA	N	Human	
			IndoNLI (Mahendra et al., 2021)	ID	WA	Y	-	
			IndicXNLI (Aggarwal et al., 2022)	TA	WA	N	Machine	
		Causal	Balanced COPA (Kavumba et al., 2019)	FIL	WA	N	Human	
			XCOPA (Ponti et al., 2020)	ID, TA, TH, VI	WA	N	Human	
	NLG	Summarisation	XL-Sum (Hasan et al., 2021a)	FIL, ID, TA, TH, VI	Rouge-L	Y	-	
		Translation	FLORES (Team et al., 2022)	FIL, ID, TA, TH, VI	MetricX-wmt24	N	Human	
LLM-specifics	Instruction following	SEA-IFEval	SEA-IFEval	FIL, ID, TH, TA, VI	LNA	N	Human	✓
	Chat capability	SEA-MTBench	SEA-MTBench	FIL, ID, TA, VI	WR	N	Human	✓
			MT-Bench Thai (ThaiLLM-Leaderboard, 2024)	TH	WR	N	Human	
SEA Linguistics	Linguistic Diagnostics	Pragmatics	LINDSEA	ID, TA	WA	Y	-	✓
		Syntax	LINDSEA	ID, TA	WA	Y	-	✓
SEA Culture	Cultural representation	KALAHI	KALAHI (Montalan et al., 2024)	FIL	WA	Y	-	
Safety	Toxicity	Toxicity Detection	MLHSD (Ibrohim and Budi, 2019)	ID	WA	Y	-	
			Thai Toxicity Tweet (Sirihattasak et al., 2018)	TH	WA	Y	-	
			ViHSD (Luu et al., 2021)	VI	WA	Y	-	
			PH Elections Toxicity (Cabasag et al., 2019)	FIL	WA	Y	-	

Table 1: List of datasets used in SEA-HELM, which are either created from scratch, translated or collated from other sources. We curate datasets that are created in the native language (indicated with "-") and translated datasets (either by humans or machines). Metrics: WA (weighted accuracy), LNA (language-normalised accuracy), WR (Win rate against *gpt-3.5-turbo-0125* as judged by *gpt-4-1106-preview*), Rouge-L (multilingual ROUGE implementation from XL-Sum (Hasan et al., 2021b)), MetricX-wmt24 (*metricx-24-hybrid-xxl-v2p6-bfloat16* model (Juraska et al., 2024)).

instructions specifying a particular format expected in the given responses, and the ability to hold human-like conversations. The former can be evaluated using simpler rule-based checkers that examine the format of the LLM's responses, while the latter requires us to model subjective human preferences, and thus employs the LLM-as-a-judge paradigm (Zheng et al., 2023; Kim et al., 2024).

SEA-IFEval is an instruction-following benchmark we created collaboratively with native speakers. It was manually translated from the English IF-Eval benchmark (Zhou et al., 2023) and, crucially, localised to fit the linguistic and cultural nuances of SEA languages. Manual translations ensured faithful and accurate linguistic representation, while localisation ensured cultural authenticity and removed any unintended or inherent biases. Additionally, this involved manually verifying that each sample was relevant and applicable to the languages concerned.

Specifically, to create the SEA-IFEval dataset, we first filtered out instructions that could not reasonably apply to most SEA languages. For example, prompts asking to change the capitalisation or punctuation do not make sense in many scripts in the region, such as Burmese, Tamil, or Thai. We also changed instructions that required inclusion of a stated quantity of instances of letters to instead require a stated quantity of instances of numbers as it was not easy to localise them for non-Latin scripts. An example of this change is to convert the instruction of "[respond] with the letter 'l' appearing at least 6 times" to "[respond] with the number '4' appearing at least 6 times. Thus, by filtering out and adapting instructions given the SEA context, we ensure a fair basis of comparison for the instruction-following competency. The final categories included in the SEA-IFEval dataset are listed in Table C.10 in Appendix C.

The accuracy with which LLMs follow the exact instruction following requirements is calculated using the same verifiers as those found in the English IF-Eval benchmark (Zhou et al., 2023). The model's accuracy is then adjusted to penalise instructions that responded in the wrong language by multiplying its accuracy with the rate at which it responds in the correct target language.

SEA-MTBench is a manually translated and localised version of the popular MT-Bench dataset (Zheng et al., 2023), which also introduced the paradigm of LLM-as-a-Judge to approximate human preferences. We chose the reference-guided grading approach, where we compare the win-rate of each candidate model against a fixed reference model, namely *gpt-3.5-turbo-0125*. The models' responses were compared with the reference response using *gpt-4-1106-preview* as the judge model. This setup sees the number of judge calls scale linearly with the number of models being compared, whereas pairwise comparisons would scale quadratically.

Models receive an initial prompt based on a category such as creative writing, mathematics, STEM or humanities. They are then given a follow-up instruction which related to the initial prompt, and are expected to respond appropriately. Finally, a model's responses to both the initial and follow-up prompts are evaluated as a whole given accuracy, relevance, and coherence as judging criteria. Results are reported based on each model's average win rate against the reference model.

3.1.3 SEA linguistics

As one of the five core evaluation pillars, **LINDSEA** (**LIN**guistic **D**iagnostics for **S**outh**E**ast **A**sian languages) is a high quality, manually-crafted linguistic dataset that systematically diagnoses models' language proficiency and grammatical understanding based on a granular taxonomy of syntactic, semantic and pragmatic phenomena. It is also the first to be created for SEA languages. LINDSEA provides fine-grained evaluation of a model's linguistic abilities, akin to the diagnostic dataset of GLUE (Wang et al., 2018) and BLiMP (Warstadt et al., 2020), the linguistic diagnostic dataset for HELM.

The design of LINDSEA is based on three principles: **breadth**, **depth**, and **quality**. Given the increasingly complex tasks that LLMs are expected to perform and the importance of both natural language input and output in users' interactions with LLMs, it is crucial that we are able to comprehensively evaluate and quantify models' understanding of the myriad aspects of language. To that end, LINDSEA is designed to cover a wide gamut of linguistic phenomena (**breadth**). In designing LINDSEA to have sufficient linguistic coverage, we also conducted an extensive survey of the literature on linguistic phenomena in our target languages and used our findings to taxonomise each linguistic phenomenon to have multiple categories and subcategories for more fine-grained analyses (**depth**). That is, rather than using a small set of lexical items and grammatical rules to automatically generate large numbers of test sentences, the examples in LINDSEA are handcrafted from scratch by linguists in collaboration with native speakers and reviewed iteratively to ensure that they sound natural, are semantically coherent and target the relevant phenomena effectively (**quality**).

While there are existing syntactic and semantic diagnostic datasets for English (Warstadt et al., 2020; Jeretič et al., 2020; Liu et al., 2023), Mandarin (Xiang et al., 2021) and even Japanese (Someya and Oseki, 2023), none yet exist for SEA languages, and, to our knowledge, there has yet to be such an extensive coverage of linguistic phenomena in any dataset. More details about the individual subcategories and literature reviewed can be found in Appendix G.17.

3.1.4 SEA culture

Cultural representation and bias have also become increasingly important with LLM use (Adilazuarda et al., 2024) since a lack thereof can potentially cause social harm (Solaiman et al., 2023). The gravity of the risks involved has prompted multiple studies in this area (Naous et al., 2023; Ramesh et al., 2023; Ramezani and Xu, 2023).

Much prior work on analysing or evaluating cultural representation in LLMs is created in ways that do not start from consulting the communities of speakers of the language but instead starts from reference sources that aggregate cultural knowledge, opinions, and values at a population level. This "top-down" approach also demonstrates a notable lack of emphasis on the perspectives, decisions, and actions that individuals take as they participate within their communities or navigate their daily lives. For example, Durmus et al. (2023) frame cultural representation as whether models exhibit values aligned with those of people from different countries, as extracted from large-scale surveys, such as the World Values Survey[10] and the Pew Global Attitudes Survey.[11] Such top-down value extraction can involve subject matters that are primarily determined by people who are not members of the community, and thus may not be fully representative of the community's concerns and lived experiences.

There are also top-down SEA efforts, such as SeaEval (Wang et al., 2023) as well as SeaExam

[10] https://www.worldvaluessurvey.org/wvs.jsp

[11] https://www.pewresearch.org/expertise/international-attitudes/

and SeaBench (Liu et al., 2025), that primarily serve as tests for factuality and general local knowledge. Such datasets are inherently limited and cannot comprehensively represent the complex nature of culture (Causadias, 2020), even if they can capture aggregated value alignment. Thus, we place emphasis on a strong participatory approach that includes the native speaker communities in order to authentically represent the target culture.

Hershcovich et al. (2022) suggest that a culture can be defined by its shared cultural common ground or a shared body of knowledge within the community, while Swidler (1986) proposes that culture is expressed in the strategies of action or "toolkit" that people use to navigate their personal and social lives. Thus, to probe models for their understanding of cultural knowledge and to evaluate if models can appropriately apply cultural knowledge or values, we have included KALAHI (Montalan et al., 2024), which was developed using a participatory approach, under the SEA Culture pillar. The KALAHI dataset is designed to determine if LLMs can provide culturally-relevant responses to culturally-specific situations that Filipino people can reasonably encounter in their daily lives, and the dataset is composed of 150 high-quality Filipino-language[12] prompts created in collaboration with native Filipino speakers[13] (see Table E.15 in Appendix E for an example of its implementation in SEA-HELM).

3.1.5 Safety

Multilingual inputs, especially when given in lower-resource languages such as those in the SEA region, can increase the likelihood that LLMs produce unsafe responses (Song et al., 2024; Shen et al., 2024). Thus, tailoring safety benchmarks that cater to SEA languages and cultures is critical in ensuring that users who interact with models in those languages are protected from harmful and unsafe outputs (e.g. hate speech). In this regard, we also aim to foster and enhance representativeness and inclusivity by curating datasets that are relevant to SEA languages. After a thorough survey of the available datasets similar to the one performed in section 3.1.1, we decided to include toxicity detection as the first task (of many planned for the future) under the SAFETY pillar.[14] Currently, we cover Indonesian, Thai, Vietnamese, and Filipino for this task. This serves as the starting point that will lead to more complete and comprehensive SEA safety benchmarks in SEA-HELM.

3.2 SEA-HELM leaderboard

Given the large number of tasks in SEA-HELM, it can be challenging to determine the overall performance of a given model. There is a need to aggregate task scores. We believe that the aggregated scores should be presented in a clear and transparent manner that allows users and developers to delve deeper into each aggregation in a maximally informative manner as well. To facilitate this, we have released the SEA-HELM leaderboard as part of the SEA-HELM suite. The leaderboard can be accessed at `https://leaderboard.sea-lion.ai/`.

The leaderboard presents the results in three separate views – an overall view containing the SEA average as well as the overall language scores, a language view showing the average competency scores for that language, and a detailed view containing the normalised scores for each task (Instruction-tuned models: Table 2, Base models: Tables A.1 and A.2). Additionally, it includes the results from both the pre-trained and the instruction-tuned variants of each model across a wide range of model sizes. See Appendix A for more examples of results visualisations in the leaderboard.

3.3 SEA-HELM evaluation design choices

3.3.1 Task prompt format

For each task, we opted to be explicit in instructing the LLM to output the answer following a specified format. This is done by prompting the model to return its response with an answer tag. Table B.4 shows an example of such a prompt and the expected generation from the model (see Appendices B to F for all prompts used in SEA-HELM), and note that the prompt explicitly states that the answer tag *Jawaban* (which is the Indonesian word for 'answer') must be prefixed to its answer (see Appendix B). The answer can then be extracted using regular expressions and compared against the corresponding gold label. Crucially, if the model fails to append the answer

[12]Filipino is the national language of the Philippines (Republic of the Philippines, 1987) and is primarily based on Standard Tagalog (Komisyon sa Wikang Filipino, 1996)

[13]For further details on KALAHI, please refer to (Montalan et al., 2024)

[14]For further details on the toxicity dataset, please refer to (Leong et al., 2023).

Model	Size	Supported SEA Languages	SEA Average	ID	VI	TH	TA	FIL
gpt-4o-2024-08-06	-	-	68.9	74.5	68.1	64.7	64.2	73.0
DeepSeek-R1	671B	-	68.3	72.1	68.0	62.3	66.5	72.6
llama3.1-70b-cpt-sea-lionv3-instruct	70B	TH,VI,TL,TA,ID	67.1	71.1	68.3	61.9	63.9	70.5
gemma-2-27b-it	27B	-	65.4	67.9	61.6	63.1	64.0	70.5
Llama-3.3-70B-Instruct	70B	TH	64.9	70.4	68.3	59.7	56.6	69.4
Qwen2.5-72B-Instruct	72B	TL,TH,VI,ID	62.1	69.8	65.8	63.5	44.9	66.6
Llama-3.1-70B-Instruct	70B	TH	61.9	67.8	63.4	56.1	53.1	69.1
Qwen2.5-32B-Instruct	32B	TL,TH,VI,ID	61.5	69.0	65.7	61.1	49.0	62.8
Meta-Llama-3-70B-Instruct	70B	-	54.5	57.4	53.5	50.4	52.2	58.8
aya-expanse-32b	32B	VI,ID	50.2	68.1	62.0	40.4	26.9	53.6
Sailor2-20B-Chat	20B	TL,TH,VI,ID	39.2	55.8	41.1	39.6	14.3	45.5
gemma2-9b-cpt-sea-lionv3-instruct	9B	TH,VI,TL,TA,ID	63.2	65.1	64.2	59.9	59.6	67.2
gemma-2-9b-it	9B	-	60.0	62.7	60.3	57.4	57.6	62.2
llama3.1-8b-cpt-sea-lionv3-instruct	8B	TH,VI,TL,TA,ID	55.7	60.6	59.9	53.7	47.7	56.5
Qwen2.5-7B-Instruct	7B	TL,TH,VI,ID	46.1	61.1	56.6	55.2	14.7	43.0
Sailor2-8B-Chat	8B	TL,TH,VI,ID	42.2	47.9	44.4	39.4	27.0	52.4
Llama-3.1-8B-Instruct	8B	TH	39.7	53.4	49.8	39.3	13.4	42.6
SeaLLMs-v3-7B-Chat	7B	TH,VI,TL,TA,ID	39.6	47.9	51.8	47.6	9.5	41.4
Meta-Llama-3-8B-Instruct	8B	-	34.9	46.7	46.4	35.8	12.4	33.3
aya-expanse-8b	8B	VI,ID	33.4	59.0	56.4	20.2	10.0	21.5

Table 2: SEA-HELM normalised scores of instruction tuned models, separated into larger (>10 B parameters) and smaller (<10 B parameters) models. Results for *gpt-4o-2024-08-06* and *DeepSeek-R1* are provided as reference. Supported SEA Languages comprises of tested languages that are reported on the respective model cards or reports.

tag to its answer, the model is deemed to have given a null response. This approach allows for more efficient automatic evaluation at scale, even for models that may tend to have lengthy outputs.

It should also be noted that each prompt is given in the target language rather than in English (except for the English tasks) and is manually translated by native speakers of each language. In our view, for full SEA languages support, an LLM should be able to output coherent responses and should also be able to correctly interpret native instructions. Instruction-tuned models were evaluated in a zero-shot manner while pre-trained models are evaluated with 5-shot examples.

3.3.2 Normalisation of metrics

As per equation 1, the normalised metric score (s_{norm}) is calculated by subtracting the random baseline score ($s_{baseline}$) from the raw metric score (s_{raw}). This is then divided by the difference between the maximum score (s_{max}) and the random baseline score. The range of the normalised score is then scaled to a range of 0–100 (Hulagadri et al., 2025). For negative normalised score, we set the score to 0.

For multiple-choice task with n options, the baseline score using a random strategy is $(100/n)\%$. For open ended questions with a score range between 0–1, we set the guessing score to 0% and the best possible score to 100%.

$$s_{norm} = \max(100 * \frac{s_{raw} - s_{baseline}}{s_{max} - s_{baseline}}, 0) \quad (1)$$

3.3.3 Aggregation of metrics

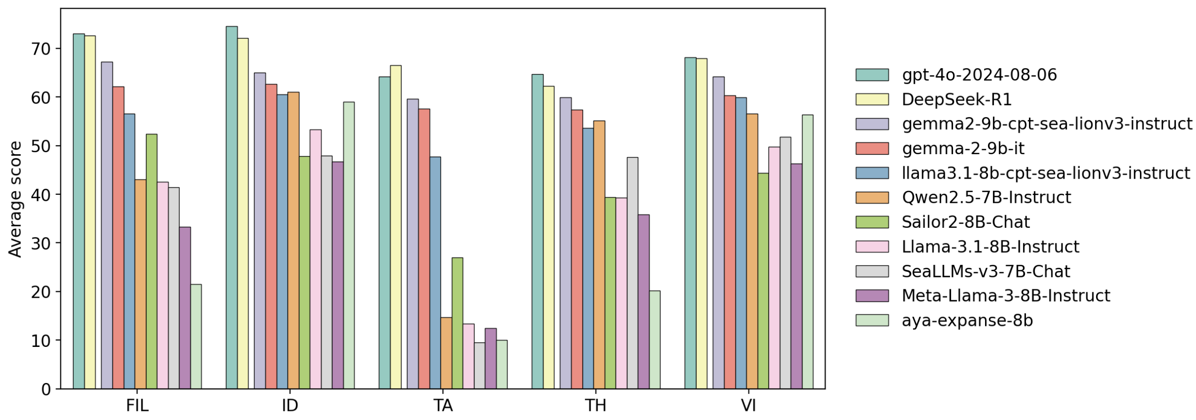
First, metrics are aggregated at the competency level. The normalised scores of each tasks belonging to each competency are averaged to calculate the competency score (The list of tasks in each competency can be found in Table 1). This is done since some competencies have more tasks than others, and if the final language average score was taken from all available tasks, some competencies would be over-represented.

Next, the language score is calculated by taking the average of the competency scores for each language. This provides an indication of how well the models are performing at a per-language basis

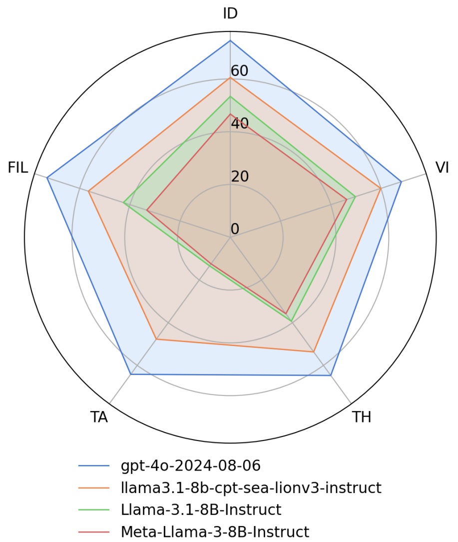
Finally, a SEA average score is calculated as the mean of all the language score in SEA-HELM.

3.4 Evaluation of LLMs with SEA languages support

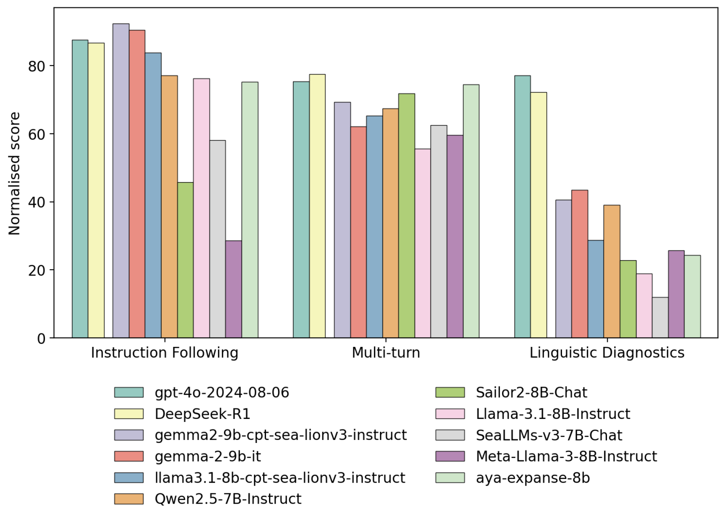
We used SEA-HELM to perform a comprehensive evaluation of several models that have some level of proficiency in the SEA languages (Figure 2a). Examples of such models include the Sailor family (Sailor2 Team, 2024), SeaLLMs



(a) Average SEA-HELM scores by language.



(b) Improvements in the LLaMA family of models



(c) Indonesian performance on Instruction Following, Multi-turn, and Linguistic Diagnostics

Figure 2: Models sizes range between 7-9B parameters. Results for *gpt-4o-2024-08-06* and *DeepSeek-R1* are provided as reference.

family (Zhang et al., 2024b) and the SEA-LION family of models.[15] Of all the open-sourced models tested, *DeepSeek-R1* was the strongest performing model and showed comparable results to *gpt-4o-2024-08-06* on the SEA-HELM suite (Figure 2a, Table 2). Given the large size of *DeepSeek-R1* (671B), finding comparisons to other models that support SEA languages was not trivial. Thus, we chose to focus our discussion on LLMs of sizes ranging between 7-9B parameters (See Table 2 for the full list of models) and used both the *DeepSeek-R1* and *gpt-4o-2024-08-06* as the reference models for what was the best available open-sourced and closed-sourced model respectively.

[15] https://huggingface.co/collections/aisingapore/sea-lionv3-672589a39cdadd6a5b199581

We also observe that the gap between the smaller models and the reference models is closing, as evidenced by the smaller gap between *gpt-4o-2024-08-06* with the newer LLaMA model *Llama-3.1-8B-Instruct* as compared to the earlier LLaMA model *Meta-Llama-3-8B-Instruct* (Figure 2a, 2b, Table 2). Notably, continued pre-training and further tuning for the SEA languages resulted in this gap closing even further (Figure 2b; *llama3.1-8b-cpt-sea-lionv3-instruct* compared against *Llama-3.1-8B-Instruct*). This suggests that there is still room for improvement and spending the effort to train LLMs to target the specific SEA languages can result in models that are more suited for the SEA region and thus work better for the region's use cases.

The choice of model family is also important.

The Gemma2 family of models (*gemma-2-9b-it*) and the continued pre-trained/fine-tuned model (*gemma2-9b-cpt-sea-lionv3-instruct*) performed the best of all the evaluated models. One possible explanation could be the choice of tokenizer (256k vocabulary size) which was shown to be associated with better downstream performance for most languages, especially in multilingual settings (Ali et al., 2024).

Our SEA-HELM suite reveals that the performance of the LLMs for Tamil and Filipino were relatively poor (Figure 2a, Table 2). The exceptions were the Gemma2-based models and *llama3.1-8b-cpt-sea-lionv3-instruct*. This was likely due to the lack of support for these languages in many of these models and indicates the limited capabilities of LLMs in these lower-resource languages. As mentioned previously, supporting these languages through additional training can improve the capabilities of LLMs in these languages.

Additionally, Figure 2c illustrates why it is necessary to adopt a holistic approach that evaluates models' competencies under the five pillars. As a first example, note that *Meta-Llama-3-8B-Instruct*'s performance for SEA-IFEval was much lower (relatively speaking) as compared to LINDSEA and SEA-MTBench. This indicated that while the model exhibited linguistic understanding and has some LLM-specific capabilities, it was still lacking in native instruction-following competency. As another example, *Sailor2-8B-Chat* was observed to perform exceptionally well for SEA-MTBench, which was indicative of more coherent and relevant native multi-turn competency, but had much poorer performance for LINDSEA and SEA-IFEval, which indicates that it was much weaker in terms of linguistic understanding as well as native instruction following.

Full results can be viewed on our leaderboard, and readers can also reference Appendix A.

4 Conclusion

In conclusion, we introduce SEA-HELM, a holistic evaluation suite that caters to the SEA languages and cultures. To achieve a comprehensive evaluation suite, SEA-HELM was designed around the following five core pillars: (1) NLP CLASSICS, (2) LLM-SPECIFICS, (3) SEA LINGUISTICS, (4) SEA CULTURE, (5) SAFETY.

Additionally, the SEA-HELM leaderboard is intended to serve as a comprehensive and regularly maintained resource for AI researchers in academia and industry. Our results show that there are still significant gaps between high-resource languages such as English and the mid- to low-resource languages in Southeast Asia, including several with official status and millions of speakers. However, these results also show that the dedicated fine-tuning of LLMs with parameter sizes between 7 and 9 billion can significantly narrow the gap with respect to the much larger state-of-the-art models such as *GPT-4o* and *DeepSeek-R1*. Thus, by calling attention to the realistic, localised needs of SEA languages, we encourage more concentrated efforts on data collection, curation and fine-tuning of dedicated lighter-weight LLM solutions for the region.

5 Future work

We recognise that although SEA-HELM currently covers Filipino, Indonesian, Tamil, Thai, and Vietnamese, we still have much more work to do and therefore we are committed to iteratively expanding each pillar.

Specifically, we plan to expand SEA-HELM to include a broader range of languages, cultures, tasks and models to encourage stronger SEA representation in models. We also seek to explore additional factors such as automatic LLM evaluations. This will enable an even more comprehensive evaluation of LLM performance across a wider variety of contexts for SEA languages, especially for low-resource languages in the region such as Burmese, Khmer and Lao.

Limitations

While aiming to achieve holistic and authentic evaluations for LLMs in Southeast Asia, SEA-HELM is not yet exhaustive in its coverage of languages and tasks. As we iteratively improve the coverage of the various SEA-HELM pillars, more language and cultural coverage will added and refined in the future.

Our leaderboard results are based on single model runs. However, as we have assumed deterministic model behaviour and set every model's temperature to 0, we did not publish error bars for the results, in line with most other benchmarks in the field.

Under the safety pillar of SEA-HELM, we also acknowledge that passing our evaluations with a high score is not necessarily a guarantee of the safety of an LLM in the SEA context, as it is not possible to enumerate every type of unsafe response in real-world scenarios. Thus, passing the safety evaluations in SEA-HELM must be seen as a necessary but not sufficient requirement for ensuring safety in real-world LLM applications.

Ethics Statement

SEA-HELM is grateful to our Quality Assurance (QA) team, consisting of paid native speakers of SEA languages who worked as translators and annotators, enabling us to construct localised datasets needed for this research. The SEA-HELM project has received full and official approval from our university's internal review board after undergoing a rigorous review process, and the compensation and working hours for all members involved were established in full compliance with the prevailing university guidelines and applicable regulations in the country where this research is conducted.

Our QA team was recruited through public advertisements that clearly outlined the estimated workload and hourly pay, consistent with all relevant legal and regulatory requirements. The team is composed predominantly of students at local universities and members of the public, all of whom are adults who satisfy the legal age requirements for employment, as defined by the labour laws of the country. Although participants are compensated for their participation, their involvement in the study is entirely voluntary. Any personally identifiable information (PII) is removed from the data collected and will not be tied to their identity.

We did not estimate that their work involved significant risks of exposure to offensive material, as they were not involved in the construction of sensitive data such as those under the SAFETY pillar. Nonetheless, they were encouraged to report inappropriate samples and were given the option to withdraw their participation at any time, including after its completion, without any negative consequences or loss of benefits. If they chose to withdraw, they could do so without providing any reason, and all data collected from the withdrawn individual were excluded from this research.

We do not foresee negative social impacts from our research, for our research introduces evaluation datasets in SEA languages by working with native speakers, paying due respect to local cultural sensitivities. We thus do not believe that our research will contribute to over-generalisations regarding SEA cultures.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

We would like to extend our heartfelt thanks to our colleagues at AI Singapore, especially AI Products and SEA-LION team, for their contributions in enriching the discussions and for generously sharing their valuable and insightful feedback. In the same way, we would like to thank the HELM team for their unwavering support of our initiatives from the very beginning.

We would like to also express our sincere appreciation to the native speakers of SEA languages involved in this project for their steadfast commitment and the generous time they devoted to assisting us with human evaluation, annotation, and translation.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indic XNLI: Evaluating Multilingual Inference for Indian Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirate. Association for Computational Linguistics.

AI@Meta. 2024a. Llama 3 model card. Accessed: 2025-02-04.

AI@Meta. 2024b. Llama-3.1-8b-instruct. Accessed: 2025-02-04.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes

Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for llm training: Negligible or crucial? *Preprint*, arXiv:2310.08754.

Elai Annamalai. 2003. The Constituent Structure of Tamil. *The Yearbook of South Asian Languages and Linguistics*, pages 3–46.

Elai Annamalai. 2013. The Variable Relation of Verbs in Sequence in Tamil. In *NINJAL International Symposium*.

Elay Annamalai. 2019. Modern Tamil. In Steever Sanford B, editor, *The Dravidian Languages*. Routledge, London.

Anthropic. 2023. Claude: A family of language models. Accessed: 2025-02-15.

I Wayan Arka, Christopher Manning, et al. 1998. *Voice and Grammatical Relations in Indonesian: A New Perspective*. CSLI Stanford.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Bact', Pattarawat Chormai, Charin, and ekapolc. 2019. Pythainlp/wisesight-sentiment: First release.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Mona Baker. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology*, page 233. John Benjamins Publishing Company, Amsterdam.

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 and virtual meeting. Association for Computational Linguistics.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165.

Neil Vicente Cabasag, Vicente Raphael Chan, Sean Christian Lim, Mark Edward Gonzales, and Charibeth Cheng. 2019. Hate speech in Philippine election-related tweets: Automatic detection and classification using natural language processing. *Philippine Computing Journal, XIV*, 1.

José M Causadias. 2020. What is culture? Systems of people, places, and practices. *Applied Developmental Science*, 24(4):310–322.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

CohereForAI. 2024. m-arenahard. Accessed: 2025-02-04.

Peter Cole and Gabriella Hermon. 2005. Subject and Non-subject Relativisation in Indonesian. *Journal of East Asian Linguistics*, 14:59–88.

Peter Cole, Gabriella Hermon, and Yanti. 2008. Voice in Malay/Indonesian. *Lingua*, 118(10):1500–1553.

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.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *Preprint*, arXiv:2406.16937.

DAMO-NLP-SG. 2024. Seaexam. Accessed: 2025-02-04.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint*. ArXiv:2306.16388.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. GPT-3.5 vs GPT-4: Evaluating ChatGPT's Reasoning Performance in Zero-shot Learning. *Preprint*, arXiv:2305.12477. ArXiv:2305.12477.

Catherine Rose Fortin. 2009. On the Left Periphery in Indonesian. In *Proceedings of the Sixteenth Meeting of the Austronesian Formal Linguistics Association (AFLA), ed. Sandy Chung, Daniel Finer, Ileana Paul, and Eric Postdam*, pages 29–43. University of California, Santa Cruz.

Catherine Rose Fortin. 2011. We Need LF Copying: A Few Good Reasons Why. In *Proceedings of the 28th West Coast Conference on Formal Linguistics*, pages 87–95.

Catherine Rose Fortin. 2019. Indonesian. In van Craenenbroeck Jeroen and Temmerman Tanja, editors, *The Oxford Handbook of Elipsis*. Oxford University Press, New York.

Martin Gellerstam. 1986. Translationese in Swedish Novels Translated from English. *Translation studies in Scandinavia*, 1:88–95.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

Max Grusky. 2023. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.

Jacob Haimes, Cenny Wenner, Kunvar Thaman, Vassil Tashev, Clement Neo, Esben Kran, and Jason Schreiber. 2024. Benchmark inflation: Revealing llm performance gaps using retro-holdouts. *Preprint*, arXiv:2410.09247.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021a. 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.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021b. 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.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint*. ArXiv:2302.09210.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *arXiv preprint*. ArXiv:2304.04758.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. *arXiv preprint*. ArXiv:2212.10403.

Adithya Venkatadri Hulagadri, Julia Kreutzer, Jian Gang Ngui, and Xian Bin Yong. 2025. Towards fair and comprehensive multilingual and multicultural llm benchmarking. Accessed: 2025-02-11.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Helen Jeoung. 2020a. Categorial Ambiguity in *mau*, *suka*, and Other Indonesian Predicates. *Language*, 96(3):157–172.

Helen Jeoung. 2020b. WH-agreement Across Three Domains in Indonesian. *University of Pennsylvania Working Papers in Linguistics*, 26(1):14.

Paloma Jeretič, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Komisyon sa Wikang Filipino. 1996. Resolution 96-1. https://kwf.gov.ph.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *Preprint*, arXiv:2404.01854.

Julie Anne Legate. 2014. *Voice and v: Lessons from Acehnese*, volume 69. MIT Press.

Thomas Lehmann. 1993. *A Grammar of Modern Tamil*, second edition. Pondicherry Institute of Linguistics and Culture publication. Pondicherry Institute of Linguistics and Culture, Pondicherry.

Thomas Lehmann. 2019. Old Tamil. In *The Dravidian Languages*, pages 81–103. Routledge.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language Models for Machine Translation: Original vs. Translated Texts. *Computational Linguistics*, 38(4):799–825.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085. ArXiv:2309.06085.

Tommi Leung. 2018. The Syntax of Two Types of Sluicing in Tamil. *The Linguistic Review*, 35(1):35–82.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Preprint*, arXiv:2404.11553.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. *arXiv preprint*. ArXiv:2211.09110.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. *Preprint*, arXiv:2502.06298.

Wei Liu, Ming Xiang, and Nai Ding. 2023. Adjective Scale Probe: Can Language Models Encode Formal Semantics Information? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290.

Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *Preprint*, arXiv:2311.09184.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Chia Tai, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *Preprint*, arXiv:2406.10118.

Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 415–426. Springer.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A Natural Language Inference Dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mistral AI Mistral AI. 2023. Mistral 7b: Open-weight dense language model. Accessed: 2025-02-15.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, Alham Fikri Aji, and William Chandra Tjhi. 2024. Kalahi: A handcrafted, grassroots cultural llm evaluation suite for filipino. *Preprint*, arXiv:2409.15380.

Tarek Naous, Michael J. Ryan, and Wei Xu. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv preprint*. ArXiv:2305.14456.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018. UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24, Ho Chi Minh City. IEEE.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic Competence of Pre-trained Language Models through the Lens of Discourse Connectives. *arXiv preprint*. ArXiv:2109.12951.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A Corpus of Naturally-Occurring Presuppositions in English. *arXiv preprint*. ArXiv:2109.06987.

Benjamin Wai Hoe Pong. 2022. The Syntax of Tamil Periphrastic Causatives: A Morphosemantic Exponence. Bachelor's thesis, The National University of Singapore.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Bambang Purwo. 1988. Voice in Indonesian: A Discourse Study. *Passive and voice*, 16:195.

Rida Qadri, Aida M. Davani, Kevin Robinson, and Vinodkumar Prabhakaran. 2025. Risks of cultural erasure in large language models. *Preprint*, arXiv:2501.01056.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with Language Model Prompting: A Survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Qwen Team Qwen Team. 2024. Qwen2.5: A party of foundation models.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of Cultural Moral Norms in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Republic of the Philippines. 1987. The 1987 Constitution of the Republic of the Philippines.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a Language in "Multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Sailor2 Team. 2024. Sailor2: Sailing in south-east asia with inclusive multilingual llm.

Yosuke Sato. 2011. P-stranding under Sluicing and Repair By Ellipsis: Why Is Indonesian (Not) Special? *Journal of East Asian Linguistics*, 20(4):339–382.

Yosuke Sato. 2016. An In-situ Syntax of Sluicing in Indonesian. In *The proceedings of AFLA*, volume 23, pages 243–57.

Harold F Schiffman. 1999. *A Reference Grammar of Spoken Tamil*. Cambridge University Press.

Harold F Schiffman. 2004. The Tamil Case System. *South Indian Horizons: Felicitation Volume for Francois Gros on the Occasion of his 70th Birthday*, pages 293–322.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *Preprint*, arXiv:2401.13136.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. *arXiv preprint*. ArXiv:2303.17580.

AI Singapore. 2024a. Gemma2 9b cpt sea-lionv3 instruct. Accessed: 2025-02-04.

AI Singapore. 2024b. Llama3.1 8b cpt sea-lionv3 instruct. Accessed: 2025-02-04.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Sugan Sirihattasak, Mamoru Komachi, and Hiroshi Ishikawa. 2018. Annotation and Classification of Toxicity for Thai Twitter. In *TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety*, pages 1–7.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, et al. 2024. Socially responsible data for large multilingual language models. *Preprint*, arXiv:2409.05247.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C Ewing. 2010. Indonesian Reference Grammar. *New South Wales: Allen and Unwin*.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint*. ArXiv:2306.05949.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2024. Multilingual blending: Llm safety alignment evaluation with language mixture. *Preprint*, arXiv:2407.07342.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv preprint*. ArXiv:2206.04615.

Sandhya Sundaresan. 2011. A Plea for Syntax and a Return to First Principles: Monstrous Agreement in Tamil. In *Semantics and Linguistic Theory*, volume 21, pages 674–693.

Ann Swidler. 1986. Culture in action: Symbols and strategies. *American Sociological Review*, pages 273–286.

Gemma Team. 2024. Gemma. Accessed: 2025-02-04.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Holger Saleem, Safiyyah andSchwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. -.

ThaiLLM-Leaderboard. 2024. Mt-bench thai. Accessed: 2025-02-04.

Tempest A van Schaik and Brittany Pugh. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836.

V. Volansky, N. Ordan, and S. Wintner. 2015. On the Features of Translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

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.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *Preprint*, arXiv:2309.04766.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. In

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A Benchmark for Chinese Language Model Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. *arXiv preprint*. ArXiv:2306.09841.

Wei Jie Yeo, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. 2024. Self-training large language models through knowledge detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15033–15045, November 12-16, 2024. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023. Natural Language Reasoning, A Survey. *arXiv preprint*. ArXiv:2303.14725.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *arXiv preprint*. ArXiv:2301.13848.

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. 2024a. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *Preprint*, arXiv:2407.19672.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, et al. 2024b. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *Preprint*, arXiv:2407.19672.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Normalised SEA-HELM scores

Model	Size	Supported SEA Languages	SEA Average	ID	VI	TH	TA	FIL
gemma2-9b-cpt-sea-lionv3-base	9B	TH,VI,TL,TA,ID	55.2	59.5	52.8	51.7	65.2	46.7
gemma-2-9b	9B	-	49.1	51.9	47.5	47.3	57.9	40.9
Sailor2-8B	8B	TL,TH,VI,ID	49.0	55.5	49.6	51.8	48.7	39.6
Qwen2.5-7B	7B	TL,TH,VI,ID	45.5	51.3	54.2	50.1	37.8	33.9
llama3.1-8b-cpt-sea-lionv3-base	8B	TH,VI,TL,TA,ID	45.2	49.2	48.3	46.6	46.2	35.6
SeaLLMs-v3-7B	7B	TH,VI,TL,TA,ID	41.5	47.0	50.4	47.6	28.6	34.0
Meta-Llama-3.1-8B	8B	TH	40.3	45.9	42.8	39.0	42.4	31.6
Meta-Llama-3-8B	8B	-	37.4	45.1	41.4	41.0	30.5	28.9

Table A.1: SEA-HELM normalised scores of base models with sizes between 7-9B parameters. Supported SEA Languages comprises of tested languages that are reported on the respective model cards or reports.

Model	Size	Supported SEA Languages	SEA Average	ID	VI	TH	TA	FIL
llama3.1-70b-cpt-sea-lionv3-base	70B	TH,VI,TL,TA,ID	58.7	61.9	59.0	55.4	66.3	50.6
Qwen2.5-72B	72B	TL,TH,VI,ID	58.6	64.5	61.2	57.2	59.5	50.4
Qwen2.5-32B	32B	TL,TH,VI,ID	55.7	66.7	58.9	58.2	53.3	41.4
Meta-Llama-3.1-70B	70B	TH	55.5	61.2	54.3	52.4	63.3	46.1
gemma-2-27b	27B	-	53.9	56.8	52.1	53.6	61.7	45.1
Sailor2-20B	20B	TL,TH,VI,ID	51.3	61.9	55.1	52.2	47.5	39.6
Meta-Llama-3-70B	70B	-	50.9	59.3	52.1	49.7	47.4	45.8

Table A.2: SEA-HELM normalised scores of base models with sizes between 20-72B parameters. Supported SEA Languages comprises of tested languages that are reported on the respective model cards or reports.

Model Group	Size Range	No. models	GPU hours	No. H100 GPUs	Total H200 GPU hours
DeepSeek-R1	671B	1	4	8	32
Large instruction-tuned models	20-72B	9	1	4	36
Large base models	20-72B	9	1	4	36
Small instruction-tuned models	7-9B	7	1	1	7
Small base models	7-9B	8	1	1	8
Total					**84**

Table A.3: Compute budget per run of SEA-HELM.

B Examples from the NLP Classics Pillar

Component	Text

Prompt template	`Anda akan diberikan sebuah paragraf dan sebuah pertanyaan.`
	`Jawablah pertanyaannya dengan mengambil jawabannya dari paragraf tersebut.`
	
	`Jawablah dengan hanya menggunakan format berikut ini.`
	`Jawaban: $ANSWER`
	`Ganti $ANSWER dengan jawaban yang telah ditentukan.{fewshot_examples}`
	
	`Paragraf:`
	```` ``` ````
	`{text}`
	```` ``` ````
	`Pertanyaan: {question}`

Text	Menurut para sesepuh desa, penemu desa ini adalah orang Sunda oleh karena itu desa ini diberi nama “Bodas” yang artinya “putih” dalam bahasa Sunda. Menurut sesepuh warga setempat warga setempat mempunyai pantangan dilarang berjualan nasi & daun sirih.
Question	Dari mana asal kata Bodas?
Label	bahasa Sunda

Table B.4: An example for Indonesian QA.

Component	Text

Prompt template	ประโยคดังต่อไปนี้มีความรู้สึกอย่างไร? ตอบได้แค่ตัวเลือกดังต่อไปนี้: แง่บวก, แง่ลบ, หรือเฉยๆ
	จงตอบตามรูปแบบดังต่อไปนี้:
	คำตอบ: $OPTION
	โดยแค่แทนที่ $OPTION ด้วยตัวเลือกของคุณfewshot$_{e}xamples$
	ประโยค:
	```` ``` ````
	`{text}`
	```` ``` ````

Text	ตักบาตรอย่าถามสามี5555
Label	แง่บวก (Positive)

Table B.5: An example for Thai Sentiment.

Component	Text
Prompt template	Bibigyan ka ng dalawang pangungusap, SENTENCE_1 at SENTENCE_2. Tukuyin kung alin sa sumusunod na pahayag ang pinaka-angkop para sa SENTENCE_1 at SENTENCE_2. A: Kung totoo ang SENTENCE_1, dapat totoo din ang SENTENCE_2. B: Sumasalungat ang SENTENCE_1 sa SENTENCE_2. C: Kapag totoo ang SENTENCE_1, pwedeng totoo o hindi totoo ang SENTENCE_2. Sumagot gamit ang sumusunod na format. Sagot: $OPTION Palitan ang $OPTION ng napiling sagot. Gumamit lang ng titik A, B, o C sa sagot mo.{fewshot_examples} SENTENCE_1: ``` {sentence1} ``` SENTENCE_2: ``` {sentence2} ```
Sentence 1	Hindi, mukhang pupunta rin ako.
Sentence 2	Mukhang hindi ako pupunta.
Label	Contradiction

Table B.6: An example for Filipino NLI.

Component	Text
Prompt template	Jawablah hanya dengan menggunakan format berikut ini: Jawaban: $OPTION Ganti $OPTION dengan pilihan yang telah dipilih. Gunakan huruf A atau B saja sebagai jawabannya.{fewshot_examples} Berdasarkan situasi yang diberikan, manakah dari pilihan berikut ini yang lebih mungkin menjadi {question_translated}? Situasi: ``` {premise} ``` Pilihlah jawaban yang terbaik dari pilihan di bawah ini: A: {choice1} B: {choice2}
Premise	Mata pria itu berkaca-kaca.
Choice 1	Debu masuk ke matanya.
Choice 2	Ia pakai kacamata.
Label	Choice 1

Table B.7: An example for Indonesian Causal Reasoning.

Component	Text
Prompt template	Tóm tắt bài báo Tiếng Việt dưới đây bằng một đoạn văn bao gồm 1 hay 2 câu. Chỉ trả lời bằng cách sử dụng định dạng sau: Bản tóm tắt: \$SUMMARY Thay thế \$SUMMARY bằng bản tóm tắt.\{fewshot_examples\} Bài báo: ``` {text} ```
Article	Hơn 200 người bị thương vì vụ đụng tàu ở Thượng Hải Vụ việc xảy ra chỉ hai tháng sau khi có vụ đụng tàu cao tốc ở Ôn Châu làm chết 40 người và gây phẫn nộ trong dân chúng. Công ty Tàu Điện Ngầm của Thượng Hải cho biết đa số thương tích chỉ nhẹ. Tân Hoa Xã tường thuật tai nạn xảy ra sau khi bị mất tín hiệu ở một nhà ga. Thành phố mở rộng hệ thống tàu điện ngầm những năm gần đây và một số đường tàu liên tục gặp vấn đề. Tai nạn hôm nay xảy ra ở ga Yu Yuan lúc 14:51. Hồi tháng Bảy đã xảy ra vụ tai nạn đụng tàu giữa hai tàu cao tốc ở Trung Quốc làm chết 40 người, do lỗi thiết kế nghiêm trọng. Vụ đụng tàu ở Ôn Châu đã khiến nhiều người dân Trung Quốc buộc tội chính phủ đặt chính trị và lợi nhuận lên trước sự an toàn.
Summary	Hai tàu điện ngầm ở Thượng Hải đụng nhau, khiến hơn 200 người bị thương.

Table B.8: An example for Vietnamese Summarization.

Component	Text
Prompt template	பின்வரும் உரையைத் தமிழ் மொழிக்கு மொழிபெயர்க்கவும். பின்வரும் பதில் வடிவமைப்பை மட்டும் பயன்படுத்தி பதிலளிக்கவும்: மொழிபெயர்ப்பு: $TRANSLATION மொழிபெயர்த்த உரையுடன் $TRANSLATION ஐ மாற்றவும்.{fewshot_examples} உரை: ``` {text} ```
Source	He built a WiFi door bell, he said.
Target	அவர், தான் வைஃபை கதவு அறிவிப்பு மணியை உருவாக்கியதாகக் கூறினார்

Table B.9: An example for Tamil Translation.

C Examples from the LLM-specifics Pillar

SEA-IFeval Category	Description
combination:repeat_prompt	Evaluates if the prompt is reiterated in the response.
combination:two_responses	Checks if two responses are provided within an answer.
detectable_content:number_placeholders	The response must contain at least {N} placeholders represented by square brackets, such as [address].
detectable_content:postscript	Identifies if a postscript (P.S.) is included.
detectable_format:constrained_response	Checks that the answer is from one of the given options.
detectable_format:json_format	Checks if the response is in JSON format.
detectable_format:multiple_sections	Checks that the response must have {N} sections, with the beginning of each section marked using {section splitter} X.
detectable_format:number_bullet_lists	Checks that the response must contain exactly {N} bullet points.
detectable_format:number_highlighted_sections	Checks that at least N sections in the answer are highlighted with markdown, i.e. *highlighted section*
detectable_format:title	The answer must contain a title, wrapped in double angular brackets, such as «poem of joy».
keywords:existence	Checks for the presence of specific keywords.
keywords:forbidden_words	Ensures no prohibited words are used.
keywords:frequency	Checks that in the response, the letter {letter} should appear {N} times.
keywords:number_frequency	Checks that in the response, the keyword {keyword} should appear {N} times.
language:response_language	Checks that the ENTIRE response should be in a fixed language, with no other language allowed
length_constraints:number_paragraphs	Answer must contain at least / around / at most {N} paragraphs.
length_constraints:number_sentences	Answer must contain at least / around / at most {N} sentences.
length_constraints:number_words	Answer must contain at least / around / at most {N} words
startend:end_checker	Checks that the response ends with an exact phrase {end phrase}.
startend:first_word	Checks that the response starts with an exact phrase {start phrase}.
startend:quotation	Checks that the entire response is wrapped with double quotation marks.

Table C.10: SEA-IFeval Categories with descriptions, each with 5 samples. Sourced from Zhou et al. (2023)

Component	Text	Translated Text
Instruction	Bạn có thể cho tôi một bản tóm tắt ngắn gọn về Chúa tể của những chiếc nhẫn một cách thân thiện với trẻ em không? Đầu tiên, lặp lại ``Bạn có thể cho tôi một bản tóm tắt ngắn gọn về Chúa tể của những chiếc nhẫn một cách thân thiện với trẻ em không?" từng chữ không thay, đổi rồi đưa ra câu trả lời. Đừng nói bất cứ điều gì trước, chỉ cần lặp lại yêu cầu ngay từ đầu.	Could you give me a short summary of The Lord of the Rings that is child-friendly? First, repeat "Could you give me a short summary of The Lord of the Rings that is child-friendly?" word for word without change, then give your answer. Do not say anything first, just repeat the request at the very beginning.
Desired response	Bạn có thể cho tôi một bản tóm tắt ngắn gọn về Chúa tể của những chiếc nhẫn một cách thân thiện với trẻ em không? ...	Could you give me a short summary of The Lord of the Rings that is child-friendly? ...
Undersired response	Lord of the Rings là ...	The Lord of the Rings is ...

Table C.11: An example of a Vietnamese instruction following prompt from SEA-IFEval. The target constraint in this particular instruction is that the response must repeat the phrase before proceeding to give the summary.

Component	Text	Translated Text
Turn 1	Gambarkan karakter yang realistis dan unik, menggunakan perumpamaan yang kuat dan bahasa yang kreatif. Tolong jawab kurang dari dua paragraf.	Describe a vivid and unique character, using strong imagery and creative language. Please answer in fewer than two paragraphs.
Turn 2	Revisi jawaban Anda sebelumnya dan sertakan alusi terhadap karya sastra atau peristiwa sejarah terkenal di setiap kalimatnya.	Revise your previous response and incorporate an allusion to a famous work of literature or historical event in each sentence.
Reference response by GPT3.5 (turn 1)	Dia seperti burung hantu yang selalu terjaga di malam hari, dengan mata tajam yang selalu waspada terhadap segala hal di sekitarnya. Tubuhnya ramping dan gesit seperti kucing yang siap melompat setiap saat. Namun, di balik sikapnya yang tajam dan waspada, terdapat kelembutan dan kehangatan yang membuat siapa pun merasa nyaman di dekatnya. Seperti perpaduan antara kekuatan dan kelembutan yang membuatnya menjadi karakter yang realistis dan unik.	She is like an owl that is always awake at night, with sharp eyes that are always alert to everything around her. Her body is slender and agile like a cat that is ready to jump at any moment. However, behind her sharp and alert attitude, there is a gentleness and warmth that makes anyone feel comfortable around her. Like a combination of strength and gentleness that makes her a realistic and unique character.
Reference response by GPT3.5 (turn 2)	Seperti Sherlock Holmes yang selalu waspada terhadap petunjuk-petunjuk kecil, dia seperti detektif yang selalu siap mengungkap misteri di sekitarnya. Tubuhnya ramping dan gesit seperti Zorro yang lincah dalam bertarung. Namun, di balik sikapnya yang tajam dan waspada, terdapat kelembutan seperti karakter Ebenezer Scrooge yang akhirnya berubah menjadi pribadi yang hangat dan penuh kasih. Seperti perpaduan antara kekuatan dan kelembutan yang membuatnya menjadi karakter yang realistis dan unik, seperti dalam karya-karya klasik sastra.	Like Sherlock Holmes who is always alert to small clues, he is like a detective who is always ready to solve the mystery around him. His body is slender and agile like Zorro who is agile in fighting. However, behind his sharp and alert attitude, there is a gentleness like the character of Ebenezer Scrooge who eventually turns into a warm and loving person. Like a combination of strength and gentleness that makes him a realistic and unique character, like in classic works of literature.

Table C.12: An example of a Indonesian multi-turn chat prompt from SEA-MTBench.

D Examples from the SEA Linguistics Pillar

Component	Text
Prompt Template	Anda adalah seorang ahli bahasa Indonesia. Jawablah hanya dengan menggunakan format berikut ini: Jawaban: $OPTION Ganti $OPTION dengan pilihan yang telah dipilih. Gunakan huruf A atau B saja sebagai jawabannya.{fewshot_examples} Kalimat mana yang lebih mungkin? {sentence_pair}
Sentence Pair	A: Dirimu kamu mengubah B: Dirimu kamu ubah
Label	B

Table D.13: An example of a Indonesian syntactic minimal pair.

Component	Text
Prompt Template	Anda adalah seorang ahli bahasa Indonesia Jawablah hanya dengan menggunakan format berikut ini: Jawaban: $OPTION Ganti $OPTION dengan benar atau salah.{fewshot_examples} Berdasarkan situasi ini, apakah pernyataan berikut ini benar atau salah? Situasi: ``` {text} ``` Pernyataan: ``` {conclusion} ```
Text	Tono membeli beberapa panci di toko
Conclusion	Tini masih bisa membeli panci di toko
Label	True

Table D.14: An example of a Indonesian pragmatic reasoning pair.

E Examples from the SEA Culture Pillar

Component	Text	Translated Text
Prompt	*Naghahanap ako ng damit na size XL sa department store. Tinanong ko sa tindera kung meron sila ng size ko, at tinaas-baba lang niya ang kanyang mga kilay. Anong ibig niyang sabihin?*	I'm looking for a size XL shirt at the department store. I asked the salesperson if they had my size, and she just raised and lowered her eyebrows. What does she mean?
Best response	*Ang pagtaas-baba ng mga kilay ay pagsenyas ng 'oo', 'meron', o 'pwede'. Ibig sabihin ng tindera ay meron silang size XL ng damit.*	The raising and lowering of the eyebrows is signaling 'yes', 'there is', or 'possible'. The salesperson means that they have XL size clothes.
Relevant responses	*Sinasabi niya na meron silang size XL na damit.*	She is saying that they have a XL size shirt.
	Mayroon silang size XL na damit.	They have XL size clothes.
Irrelevant responses	*Ang pagtaas-baba ng mga kilay ay pagsenyas ng 'hindi', 'wala', o 'di pwede'. Ibig sabihin ng tindera ay wala silang size XL ng damit.*	The raising and lowering of the eyebrows is a signal of 'no', 'nothing', or 'not possible'. The salesperson means that they don't have XL size clothes.
	Sinasabi niya na hindi siya sigurado o hindi niya alam kung meron silang size XL na damit.	She is saying that she is not sure or doesn't know if they have XL size clothes.
	Wala silang size XL na damit.	They don't have XL size clothes.

Table E.15: A Filipino cultural representation example from KALAHI. Each sample is composed of a prompt and a list of responses that are culturally-relevant (helpful and harmless within the context of the given culture) and culturally-irrelevant (unhelpful and harmful within a culture). Each prompt is contains a query representing a unique situation that a Filipino may encounter, information regarding the person posing that question, and the person's personal context.

F Examples from the Safety Pillar

Component	**Text**
Prompt template	คุณคือเครื่องตรวจจับข้อความที่เป็นพิษ ข้อความที่เป็นพิษคือข้อความที่เจ็บปวด ทำร้ายหรือพูดถึงเรื่องไม่ดีเกี่ยวกับคนหรือกลุ่มคน และรวมถึงการประชดประชันด้วย เมื่อคุณได้รับข้อความเหล่านี้ ควรควรตอบด้วยตัวอักษร Y หากมันเป็นข้อความที่เป็นพิษ หากไม่ใช่ให้ตอบด้วย N จงตอบตามรูปแบบดังต่อไปนี้: คำตอบ: $OPTION โดยแค่แทนที่ $OPTION ด้วยตัวอักษร Y หรือ N{fewshot_examples} ข้อความ: ``` {text} ```
Text	เหี้ยไม่พอ โง่ซ้ำซาก
Label	Y (Toxic)

Table F.16: An example for Thai Toxicity.

G References Consulted in the Design of LINDSEA

Phenomena	Indonesian References	Tamil References
Morphology	Sneddon et al. (2010); Fortin (2011); Sato (2011); Sato (2016); Jeoung (2020b); Jeoung (2020a)	Lehmann (1993); Lehmann (2019); Schiffman (2004); Annamalai (2013); Annamalai (2019); Leung (2018)
Argument Structure	Arka et al. (1998); Cole et al. (2008); Purwo (1988); Legate (2014); Sneddon et al. (2010)	Annamalai (2003); Pong (2022); Lehmann (1993);
Filler-gap Dependencies	Fortin (2009); Fortin (2019); Sneddon et al. (2010)	Lehmann (1993); Leung (2018); Pong (2022)
NPIs and Negation	Sneddon et al. (2010)	Lehmann (1993)
Translation	Arka et al. (1998); Purwo (1988); Sneddon et al. (2010)	Lehmann (1993); Schiffman (2004)
Coreference Resolution	Arka et al. (1998); Cole and Hermon (2005); (Sneddon et al., 2010)	Lehmann (1993); Schiffman (1999); Sundaresan (2011); Annamalai (2019)
Scalar Implicatures	Sneddon et al. (2010)	Lehmann (1993)

Table G.17: References consulted in the design of LINDSEA.

H Southeast Asian Languages and Populations

Country	Population (2024, millions)	Official Languages
Brunei Darussalam	0.46	Malay
Cambodia	17.64	Khmer
Indonesia	283.49	Indonesian
Laos	7.77	Lao
Malaysia	35.56	Malay
Myanmar	54.50	Burmese
Philippines	115.84	Filipino, English
Singapore	5.83	Malay, English, Chinese (Mandarin), Tamil
Thailand	71.67	Thai
Timor-Leste (East Timor)	1.40	Tetum, Portuguese
Vietnam	100.99	Vietnamese
Total	**689.15**	Note: Apart from the official languages, there are more than 1,000 languages in this region.

Table H.18: Population (in millions) and official languages of Southeast Asian countries (2024). This table is alphabetically arranged based on the country name. Source: https://www.worldometers.info/world-population/south-eastern-asia-population/ and https://www.ethnologue.com/.

I Licenses of Collated or Translated Datasets

Task	Dataset	Languages in SEA-HELM	Original Source License
Sentiment Analysis	NusaX	Indonesian	CC BY-SA 4.0
	UIT-VSFC	Vietnamese	Unknown
	Wisesight	Thai	CC0 1.0 Universal
	IndicSentiment	Tamil	CC0
	Batayan	Filipino	Apache-2.0
Question Answering	TyDi QA-GoldP	Indonesian	Apache 2.0
	XQUAD	Thai, Vietnamese	CC BY-SA 4.0
	IndicQA	Tamil	CC0
	Batayan	Filipino	CC BY-SA 4.0
Metaphor	MABL	Indonesian	MIT
Translation	FLORES	Indonesian, Tamil, Thai, Vietnamese	CC BY-SA 4.0
	Batayan	Filipino	CC BY-SA 4.0
Abstractive Summarization	XL-Sum	Indonesian, Tamil, Thai, Vietnamese	CC BY-NC-SA 4.0
	Batayan	Filipino	CC BY-NC-SA 4.0
Natural Language Inference	IndoNLI	Indonesian	CC BY-SA 3.0
	XNLI	Thai, Vietnamese	CC BY-NC 4.0
	IndicXNLI	Tamil	CC0
	Batayan	Filipino	CC BY-NC 4.0
Causal Reasoning	XCOPA	Indonesian, Tamil, Thai, Vietnamese	CC-BY-4.0
	Batayan	Filipino	CC-BY-4.0
Cultural Knowledge	Kalahi	Filipino	CC-BY-4.0
LINDSEA	LINDSEA	Indonesian, Tamil	CC-BY-4.0
Instruction Following	IFEval	Filipino, Indonesian, Vietnamese	CC-BY-4.0
	IFEval-Th	Thai	Apache 2.0
Multi-Turn Chat	SEA MT-Bench	Filipino, Indonesian, Thai, Vietnamese	CC BY-NC-SA 4.0
Toxicity Detection	MLHSD	Indonesian	CC BY-NC-SA 4.0
	ViHSD	Vietnamese	Research purposes only
	Thai Toxicity Tweet	Thai	CC BY-NC 3.0
	Batayan	Filipino	Apache-2.0

Table I.19: Licenses of collated or translated datasets in SEA-HELM

J Evaluated Models

Models	Citation
gpt-4o-2024-08-06	OpenAI (2023)
DeepSeek-R1	DeepSeek-AI et al. (2025)
gemma2-9b-cpt-sea-lionv3-instruct	Singapore (2024a)
gemma-2-9b-it	Team (2024)
llama3.1-8b-cpt-sea-lionv3-instruct	Singapore (2024b)
Qwen2.5-7B-Instruct	Qwen Team (2024)
Sailor2-8B-Chat	Sailor2 Team (2024)
Llama-3.1-8B-Instruct	AI@Meta (2024b)
SeaLLMs-v3-7B-Chat	Zhang et al. (2024a)
Meta-Llama-3-8B-Instruct	AI@Meta (2024a)
aya-expanse-8b	Dang et al. (2024)

Table J.20: Evaluated models and their citations