SEA-HELM：东南亚语言模型综合评估

> **译注：** 本文翻译自论文 "Southeast Asian Holistic Evaluation of Language Models"（arXiv:2502.14301），原作者为 Yosephine Susanto 等人（AI Singapore / 新加坡国立大学 / 斯坦福大学 CRFM）。技术术语首次出现时保留英文并附中文说明。参考文献保留英文原文。
>
> 原文链接：https://arxiv.org/abs/2502.14301

作者： Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, William Chandra Tjhi

机构： AI Singapore（新加坡人工智能机构）、新加坡国立大学、斯坦福大学基础模型研究中心（CRFM）

摘要

随着大语言模型（Large Language Models, LLMs）新能力的快速涌现，对严格的多语言和多文化综合性基准测试（benchmark）的需求日益迫切。尽管现有的 LLM 基准测试能够评估 LLMs 在英语以及各种中低资源语言（包括东南亚地区语言）中的特定能力，但迄今为止尚未开发出针对东南亚语言的全面且真实的评估套件。

在此，我们提出 **SEA-HELM**（前身为 BHASA），这是一个强调东南亚语言的综合性语言和文化 LLM 评估套件，包含五大核心支柱：（1）经典 NLP（NLP Classics）；（2）LLM 专项

（LLM-specifics）；（3）东南亚语言学（SEA Linguistics）；（4）东南亚文化（SEA Culture）；（5）安全性（Safety）。

SEA-HELM 目前支持菲律宾语、印度尼西亚语、泰米尔语、泰语和越南语。我们还推出了 SEA-HELM 排行榜（https://leaderboard.sea-lion.ai/），使用户能够系统化且友好地了解模型的多语言和多文化表现。

1 引言

生成式自然语言处理（NLP）方法通过大语言模型的普及，使得许多传统的 NLP 评估数据集已被攻破（Haimes et al., 2024）、过时或饱和（Liu et al., 2024）。虽然 LLMs 本质上是训练来预测序列中下一个 token（词元）的，但它们展现了显著的涌现能力（emergent capabilities），包括摘要、问答、翻译、编码和高级推理（Brown et al., 2020; Yeo et al., 2024）。它们还被广泛用于新应用，例如能够进行持续开放式对话的聊天机器人（Dam et al., 2024）。这一进步导致 LLM 能力范围与严格评估它们的数据集和框架之间存在显著差距。传统的 NLP 评估方法强调与预定义的标准答案对齐，不足以衡量 LLMs 的复杂能力。这种差距在东南亚的低资源语言中更为严重，因为互联网上的训练和测试数据都非常匮乏（Li et al., 2024）。

图 1： SEA-HELM 的五大评估支柱，构成我们综合整合的评估方法。

东南亚地区拥有近 7 亿使用者，涵盖超过 1000 种语言。虽然该地区代表了全球近 10% 的人口，约占世界语言总数的七分之一，但大多数这些语言仍未被主流 LLMs 如 Mistral（Mistral AI, 2023）和 Claude（Anthropic, 2023）所支持。数据匮乏以及数字访问和代表性不均等造成的复杂性，阻碍了该地区这些语言 LLMs 的发展。一些东南亚语言使用非拉丁文字书写，这更增加了分词器（tokenizer）在处理有限数据时的挑战。

尽管存在上述障碍，东南亚的多语言 LLM 和基准测试开发仍在努力缩小差距并适应该领域的当前趋势。一些模型现在明确支持东南亚语言，并声称提供东南亚文化知识的代表性（Sailor2 Team, 2024; Zhang et al., 2024b; Bai et al., 2023; Dang et al., 2024）。也有许多基准测试声称衡量 LLMs 在东南亚地区的多语言和多文化能力（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）。然而，目前尚未出现针对东南亚文化和语言能力评估 LLMs 的综合性、真实性基准测试套件。

因此，我们开发了 **SEA-HELM**（SouthEast Asian Holistic Evaluation of Language Models，东南亚语言模型综合评估），一个系统化、集成化且持续维护的基准测试套件，旨在以针对性和全面的方式衡量 LLMs 的东南亚语言和文化能力。SEA-HELM 通过整合本地化评估数

据集和 LLM 提示词（prompt）、对模型统一运行测试以实现标准化比较、按语言、任务和模型汇总呈现结果来实现集成化。我们认为没有单一指标能够说明模型对东南亚的适用性，因此 SEA-HELM 被设计为测试一套综合的能力集（如图 1 所示）。

具体而言，SEA-HELM 组织为五大评估支柱：（1）经典 NLP；（2）LLM 专项；（3）东南亚语言学；（4）东南亚文化；（5）安全性，这五大支柱共同涵盖了每种东南亚语言的广泛任务范围，确保从语言细微差别到文化代表性的各个相关方面都得到考虑。这五大支柱也经过精心严格的构建，以实现该地区 LLMs 的公平、透明和真实的多语言多文化评估。我们刻意纳入社区参与，在数据集规划和构建的每个阶段让东南亚语言的母语者参与其中，以确保语言准确性和文化真实性。

SEA-HELM 的贡献总结如下：

- SEA-HELM 是一个精选的东南亚数据集套件，这些数据集会一起评估，结果将在公开可见的排行榜上展示。
- SEA-HELM（a）将现有的英语安全和 NLP 任务原生适配并翻译为菲律宾语、印度尼西亚语、泰米尔语、泰语和越南语；（b）创建了人工翻译的 SEA-IFEval 和 SEA-MTBench 数据集；（c）创建了跨任务和语言一致的 LLM 提示词模板。
- SEA-HELM 包括我们为印度尼西亚语和泰米尔语开发的细粒度语言诊断（LINDSEA）新数据集。SEA-HELM 还包括与菲律宾社区成员合作开发的菲律宾文化评估数据集 Kalahi（Montalan et al., 2024）。

综合 SEA-HELM 中包含的其他数据集，我们相信这些使 SEA-HELM 成为迄今为止对东南亚语言准确且真实的评估套件。它可以作为未来扩展覆盖其他东南亚语言（如高棉语、老挝语、缅甸语等）的基础，我们将在下一次迭代中探索。

2 相关工作

2.1 LLM 评估

多年来，AI 从业者采用单一任务或更罕见的综合方法来评估 LLMs 的性能和能力。评估 LLMs 的常见任务包括翻译（Hendy et al., 2023）、摘要（Zhang et al., 2023）、决策制定（Shen et al., 2023）、检测标量蕴含（scalar implicatures，即通过对话推断隐含的量级信息）（Jeretič et al., 2020; Pandia et al., 2021; Hu et al., 2023; Liu et al., 2023）以及预设（presuppositions）（Jeretič et al., 2020; Parrish et al., 2021）。此外，语言学评估

（Warstadt et al., 2020; Xiang et al., 2021; Someya and Oseki, 2023）和文化代表性评估（Durmus et al., 2023; Atari et al., 2023）也日益被认为是评估语言模型效能和公平性的重要标准。

在综合方法方面，斯坦福大学推出了 **HELM**（Holistic Evaluation of Language Models，语言模型综合评估）（Liang et al., 2022），旨在跨广泛任务评估 LLMs，包括语言能力、推理、知识、记忆、虚假信息、偏见和毒性。Google 推出了 **BIG-Bench**（Srivastava et al., 2023），这是一个众包计划。类似地，OpenAI 推出了 **OpenAI Evals**，一个邀请用户创建自定义评估数据集的众包系统。

2.2 东南亚语言的 LLM 评估

近年来，越来越多的关注被投向英语以外的 LLM 训练和评估。有越来越多的工作（Sailor2 Team, 2024; Zhang et al., 2024b; Bai et al., 2023; Dang et al., 2024）评估 LLMs 在东南亚语言中各种任务的表现。其中大多数也试图涵盖广泛的语言（如印度尼西亚语、泰语、菲律宾语）。为了实现如此大的语言覆盖范围，通常使用机器翻译和合成生成来生成多语言基准数据集。

然而，使用机器翻译和合成生成的基准测试，且很少有社区参与，引发了对其真实性和可靠性的质疑。自动翻译经常遗漏目标语言中固有的文化细微差异，并可能导致翻译错误和偏见（Singh et al., 2024）。这可能导致文化擦除（cultural erasure），进一步加深刻板或缺乏多样性的观点（Qadri et al., 2025）。因此，有必要开发真实的、经人工验证的多语言评估数据集和指标。Singh et al.（2024）、Romero et al.（2024）和 Koto et al.（2024）等工作通过采用参与式框架（participatory framework）解决了上述问题（Birhane et al., 2022; Smart et al., 2024）。参与式框架也是 SEA-HELM 设计理念的核心，因为它确保了语言准确性和文化真实性。

3 SEA-HELM

为了解决东南亚地区缺乏综合性多语言多文化评估的问题，我们设计并开发了 SEA-HELM，其灵感来源于 HELM（Liang et al., 2022）。该评估套件由五大核心支柱组成：（1）经典 NLP；（2）LLM 专项；（3）东南亚语言学；（4）东南亚文化；（5）安全性，并已与 HELM 集成。任务和语言的分布详见表 1。SEA-HELM 目前支持五种东南亚语言——菲律宾语、印度尼西亚语、泰米尔语、泰语和越南语，使用户和 AI 从业者能够评估 LLMs 在这些语言上的整体表现。

表 1：SEA-HELM 使用的数据集列表

支柱	能力	任务	数据集	语言	指标	原生/翻译
经典 NLP	自然语言理解 (NLU)	情感分析	PH Elections Sentiment	FIL	WA	原生
			NusaX	ID	WA	原生
			IndicSentiment	TA	WA	人工翻译
			Wisesight	TH	WA	原生
			UIT-VSFC	VI	WA	原生
		问答 (QA)	TyDi QA-GoldP	ID	F1	原生
			IndicQA	TA	F1	原生
			XQUAD	TH, VI	F1	人工翻译
		多选题 QA	Belebele	FIL	F1	人工翻译
		隐喻	MABL	ID	WA	原生
	自然语言推理 (NLR)	NLI	XNLI	FIL, TH, VI	WA	人工翻译
			IndoNLI	ID	WA	原生
			IndicXNLI	TA	WA	机器翻译
		因果推理	Balanced COPA	FIL	WA	人工翻译
			XCOPA	ID, TA, TH, VI	WA	人工翻译
	自然语言生成 (NLG)	摘要	XL-Sum		Rouge-L	原生

支柱	能力	任务	数据集	语言	指标	原生/翻译
				FIL, ID, TA, TH, VI		
		翻译	FLORES	FIL, ID, TA, TH, VI	MetricX-wmt24	人工翻译
LLM 专项	指令遵循	SEA-IFEval	SEA-IFEval	FIL, ID, TH, TA, VI	LNA	人工翻译
	对话能力	SEA-MTBench	SEA-MTBench	FIL, ID, TA, VI	WR	人工翻译
			MT-Bench Thai	TH	WR	人工翻译
东南亚语言学	语言诊断	语用学	LINDSEA	ID, TA	WA	原生
		句法	LINDSEA	ID, TA	WA	原生
东南亚文化	文化代表性	KALAHI	KALAHI	FIL	WA	原生
安全性	毒性	毒性检测	MLHSD	ID	WA	原生
			Thai Toxicity Tweet	TH	WA	原生
			ViHSD	VI	WA	原生
			PH Elections Toxicity	FIL	WA	原生

指标说明： WA（加权准确率）、LNA（语言归一化准确率）、WR（以 gpt-3.5-turbo-0125 为参照、gpt-4-1106-preview 为评判的胜率）、Rouge-L（XL-Sum 多语言 ROUGE 实现）、MetricX-wmt24（metricx-24-hybrid-xxl-v2p6-bfloat16 模型）。

3.1 核心支柱

3.1.1 经典 NLP

首先，在**自然语言理解**（NLU）能力方面，我们纳入了问答（抽取式问答）和情感分析任务。其次，在**自然语言生成**（NLG）能力方面，我们纳入了翻译（英译本地语言和本地语言译英）和抽象式摘要任务。第三，在**自然语言推理**（NLR）能力方面，我们纳入了因果推理和自然语言推断（NLI，Natural Language Inference）任务。

我们尽可能选择由母语者用本地语言原始编写的数据集。否则，现有的英语数据集由母语者仔细翻译。这很重要，因为翻译数据集通常包含翻译腔（translationese）的元素（Gellerstam, 1986），这与母语者原始书写的文本存在显著差异（Baker, 1993; Lembersky et al., 2012; Volansky et al., 2015; Riley et al., 2020）。

3.1.2 LLM 专项

随着 LLMs 催生前所未有的 NLP 应用，有必要为这些高阶任务开发自动化的专用评估指标。SEA-HELM 聚焦于两项特定能力——遵循人类指令指定特定响应格式的能力，以及进行类人对话的能力。前者可以使用较简单的基于规则的检查器来检查 LLM 响应的格式，而后者需要建模主观的人类偏好，因此采用 **LLM-as-a-judge**（LLM 作为评判者）范式。

SEA-IFEval 是我们与母语者合作创建的指令遵循基准测试。它从英语 IF-Eval 基准测试（Zhou et al., 2023）手动翻译而来，并且关键的是，进行了本地化以适应东南亚语言的语言和文化细微差别。手动翻译确保忠实准确的语言表达，而本地化确保文化真实性并消除任何无意或固有的偏见。这包括手动验证每个样本是否与相关语言相关且适用。

具体来说，在创建 SEA-IFEval 数据集时，我们首先过滤掉不能合理适用于大多数东南亚语言的指令。例如，要求更改大小写或标点的提示在该地区的许多文字系统中没有意义，如缅甸文、泰米尔文或泰文。我们还将要求特定字母频率的指令改为要求特定数字频率，因为对于非拉丁文字，前者不容易本地化。因此，通过过滤和调整东南亚语境下的指令，我们确保了指令遵循能力比较的公平基础。模型遵循精确指令要求的准确率将被计算。然后，通过将准确率乘以模型用正确目标语言回复的比率来调整，以惩罚虽遵循指令但使用错误语言回复的情况。

SEA-MTBench 是流行的 MT-Bench 数据集（Zheng et al., 2023）的手动翻译和本地化版本，该数据集也引入了 LLM-as-a-Judge 范式来近似人类偏好。我们选择了参考引导评分方法，将每个候选模型与固定参考模型（即 gpt-3.5-turbo-0125）的胜率进行比较。模型的回复使用 gpt-4-1106-preview 作为评判模型与参考回复进行比较。这种设置使得评判调用次数随被比较模型数量线性增长，而成对比较将呈二次方增长。

模型收到基于类别（如创意写作、数学、STEM 或人文学科）的初始提示，然后获得与初始提示相关的后续指令，并被期望做出适当回应。最后，根据准确性、相关性和连贯性作为评判标准，对模型对初始和后续提示的回复进行整体评估。结果基于每个模型相对于参考模型的平均胜率报告。

3.1.3 东南亚语言学

作为五大核心评估支柱之一，**LINDSEA**（LINguistic Diagnostics for SouthEast Asian languages，东南亚语言的语言诊断）是一个高质量、手工构建的语言数据集，基于句法（syntax）、语义（semantics）和语用（pragmatics）现象的细粒度分类法，系统地诊断模型的语言能力和语法理解。这也是首个为东南亚语言创建的此类数据集。LINDSEA 提供对模型语言能力的细粒度评估，类似于 GLUE 的诊断数据集（Wang et al., 2018）和 BLiMP（Warstadt et al., 2020）——HELM 的语言诊断数据集。

LINDSEA 的设计基于三个原则：**广度、深度和质量**。鉴于 LLMs 被期望执行日益复杂的任务，以及自然语言输入输出在用户与 LLMs 交互中的重要性，我们能够全面评估和量化模型对语言众多方面的理解至关重要。为此，LINDSEA 被设计为覆盖广泛的语言现象（广度）。在设计 LINDSEA 以具有足够的语言覆盖时，我们还对目标语言中语言现象的文献进行了广泛调研，并利用研究结果将每个语言现象分类为多个类别和子类别，以进行更细粒度的分析（深度）。也就是说，LINDSEA 中的示例不是使用少量词汇和语法规则自动生成大量测试句子，而是由语言学家与母语者合作从零开始手工构建，并经过迭代审查以确保它们听起来自然、语义连贯并有效地针对相关现象（质量）。

虽然已有针对英语（Warstadt et al., 2020; Jeretic ˇet al., 2020; Liu et al., 2023）、普通话（Xiang et al., 2021）甚至日语（Someya and Oseki, 2023）的句法和语义诊断数据集，但东南亚语言尚无此类数据集，据我们所知，也尚未有任何数据集具有如此广泛的语言现象覆盖。

3.1.4 东南亚文化

随着 LLM 的使用，文化代表性和偏见问题也变得日益重要，因为缺乏代表性可能造成社会伤害（Solaiman et al., 2023）。相关风险的严重性促使了该领域的多项研究（Naous et al., 2023; Ramesh et al., 2023; Ramezani and Xu, 2023）。

此前关于分析或评估 LLMs 中文化代表性的许多工作展示了我们所称的**"自上而下"方法**，即主要依赖在人口层面汇总文化知识、观点和价值观的参考来源进行数据收集和/或数据创建的策略。这种方法在个人参与社区活动或日常生活中所采取的观点、决策和行动方面也明显缺乏关注。

例如，Durmus et al.（2023）将文化代表性评估框定为确定模型表现出的价值观是否与来自不同国家的人的价值观一致，这些价值观从大规模调查中提取，如世界价值观调查（World Values Survey）和皮尤全球态度调查（Pew Global Attitudes Survey）。这种自上而下的价值提取可能涉及主要由非目标群体成员确定的主题，因此可能无法完全代表社区的关切和生活经验。

也有自上而下的东南亚工作，如 SeaEval（Wang et al., 2023）以及 SeaExam 和 SeaBench（Liu et al., 2025），主要作为事实性和一般本地知识的测试。这种自上而下的评估数据集本质上是有限的，无法全面代表文化的多面性和复杂性（Causadias, 2020），即使它们可以捕获汇总的价值对齐。因此，我们强调采用强参与式方法，纳入母语者社区以真实地代表目标文化。

Hershcovich et al.（2022）建议文化可以通过其共享的文化共同基础（shared cultural common ground）或社区内的共享知识体系来定义，而 Swidler（1986）提出文化体现在人们用来导航个人和社会生活的行动策略或"工具箱"中。因此，为了探查模型对文化知识的理解并评估模型是否能恰当运用文化知识或价值观，我们在东南亚文化支柱下纳入了 **Kalahi**（Montalan et al., 2024），这是我们使用参与式方法开发的数据集。Kalahi 数据集旨在确定 LLMs 是否能够对菲律宾人在日常生活中合理遇到的文化特定情境提供文化相关的回应，该数据集由与菲律宾语母语者合作创建的 150 个高质量提示词组成。

3.1.5 安全性

多语言输入，尤其是使用东南亚地区低资源语言时，可能增加 LLMs 产生不安全回复的可能性（Song et al., 2024; Shen et al., 2024）。因此，定制适合东南亚语言和文化的安全基准测试对于确保以这些语言与模型交互的用户免受有害和不安全输出（如仇恨言论）的侵害至关重要。在这方面，我们也旨在通过策划与东南亚语言相关的数据集来促进和增强代表性和包容性。经过类似于第 3.1.1 节的可用数据集全面调研，我们决定将**毒性检测**作为安全支柱下的第一个任务（未来计划更多）。目前该任务覆盖印度尼西亚语、泰语、越南语和菲律宾语。这将作为起点，引向 SEA-HELM 中更完整和全面的东南亚安全基准测试。

3.2 SEA-HELM 排行榜

鉴于 SEA-HELM 中任务数量众多，确定给定模型的整体表现可能具有挑战性。需要对任务分数进行聚合。我们认为聚合分数应以清晰透明的方式呈现，使用户和开发者也能以最大信息量的方式深入了解每个聚合。为此，我们发布了 SEA-HELM 排行榜作为 SEA-HELM 套件的一部分。

排行榜以三个独立视图呈现结果——包含 SEA 平均分和各语言总分的**总体视图**、显示该语言各能力平均分的**语言视图**，以及包含每个任务归一化分数的**详细视图**。此外，它还包括每个模型的预训练（pre-trained）和指令微调（instruction-tuned）变体的结果，涵盖广泛的模型规模。

3.3 评估细节

3.3.1 任务提示词格式

对于每个任务，我们选择明确指示 LLM 按照指定格式输出答案。这是通过提示模型以答案标签返回其回复来实现的。例如，提示明确要求答案标签 "Jawaban"（印度尼西亚语中"答案"一词）必须作为其答案的前缀。然后可以使用正则表达式提取答案并与相应的标准答案进行比较。关键的是，如果模型未在其答案中附加答案标签，则该模型被视为给出了空回复。这种方法允许在大规模情况下进行更高效的自动评估，即使对于倾向于冗长输出的模型也是如此。

还应注意，每个提示词都以目标语言而非英语给出（英语任务除外），并由各语言的母语者手动翻译。在我们看来，要完全支持东南亚语言，LLM 不仅应能输出连贯的回复，还应能正确理解母语指令。指令微调模型以零样本（zero-shot）方式评估，而预训练模型以 5 样本（5-shot）方式评估。

3.3.2 指标聚合

我们选择首先将每个任务分配到第 3.1 节中定义的各种能力之一来聚合指标。然后我们对每个指标分数进行归一化，考虑随机基线分数并将其重新缩放到 0-100 的范围（Hulagadri et al., 2025）。然后对每个能力内的归一化分数取平均值。对于每种语言，我们取各能力的平均值得到总体语言分数。最后，SEA 平均分作为所有支持语言分数的均值计算。

3.4 支持东南亚语言的 LLMs 评估

图 2： 模型参数规模在 7-9B 之间。gpt-4o-2024-08-06 和 DeepSeek-R1 的结果作为参考提供。- (a) 按语言划分的 SEA-HELM 平均分 - (b) LLaMA 系列模型的改进 - (c) 印度尼西亚语在指令遵循、多轮对话和语言诊断上的表现

我们使用 SEA-HELM 对几个在东南亚语言上具有一定能力的模型进行了全面评估（图 2a）。此类模型的示例包括 Sailor 系列（Sailor2 Team, 2024）、SeaLLMs 系列（Zhang et al., 2024b）和 SEA-LION 系列模型。在所有测试的开源模型中，**DeepSeek-R1 是表现最强的模型**，在 SEA-HELM 套件上显示出与 gpt-4o-2024-08-06 相当的结果。鉴于 DeepSeek-R1 的大参数量（671B），找到支持东南亚语言的同等模型进行比较并不容易。因此，我们选择将讨论重点放在参数规模在 7-9B 之间的 LLMs 上，并使用 DeepSeek-R1 和 gpt-4o-2024-08-06 分别作为最佳可用的开源和闭源参考模型。

我们还观察到**较小模型与参考模型之间的差距正在缩小**，较新的 LLaMA 模型 Llama-3.1-8B-Instruct 与 gpt-4o-2024-08-06 之间的差距比早期的 Meta-Llama-3-8B-Instruct 更小（图 2a、2b）。值得注意的是，针对东南亚语言的持续预训练和进一步微调使这一差距进一步缩小（图 2b；llama3.1-8b-cpt-sea-lionv3-instruct 与 Llama-3.1-8B-Instruct 的比较）。这表明仍有改进空间，花费精力训练针对特定东南亚语言的 LLMs 可以产生更适合东南亚地区的模型，从而更好地服务于该地区的使用场景。

模型系列的选择也很重要。 Gemma2 系列模型（gemma-2-9b-it）及其持续预训练/微调模型（gemma2-9b-cpt-sea-lionv3-instruct）在所有评估模型中表现最佳。一个可能的解释是分词器的选择（256k 词汇量），这被证明与大多数语言的更好下游性能相关，尤其是在多语言设置中（Ali et al., 2024）。

我们的 SEA-HELM 套件揭示，**LLMs 在泰米尔语和菲律宾语上的表现相对较差**（图 2a）。例外的是基于 Gemma2 的模型和 llama3.1-8b-cpt-sea-lionv3-instruct。这很可能是由于许多模型缺乏对这些语言的支持，表明 LLMs 在这些较低资源语言中的能力有限。如前所述，通过额外训练支持这些语言可以提升 LLMs 在这些语言中的能力。

此外，图 2c 说明了为什么有必要采用综合方法在五大支柱下评估模型能力。作为第一个例子，请注意 Meta-Llama-3-8B-Instruct 在 SEA-IFEval 上的表现相对于 LINDSEA 和 SEA-MTBench 要差得多。这表明虽然该模型展示了语言理解能力并具有一定的 LLM 专项能力，但在母语指令遵循能力方面仍有不足。另一个例子是，Sailor2-8B-Chat 被观察到在 SEA-MTBench 上表现出色，这表明其具有更连贯和相关的母语多轮对话能力，但在 LINDSEA 和 SEA-IFEval 上表现差得多，这表明其在语言理解以及母语指令遵循方面要弱得多。

4 结论

总之，我们介绍了 SEA-HELM，一个面向东南亚语言和文化的综合评估套件。为实现全面的评估套件，SEA-HELM 围绕以下五大核心支柱设计：（1）经典 NLP；（2）LLM 专项；（3）东南亚语言学；（4）东南亚文化；（5）安全性。

此外，SEA-HELM 排行榜旨在作为学术界和工业界 AI 研究人员的综合且定期维护的资源。我们的结果表明，英语等高资源语言与东南亚中低资源语言之间仍然存在显著差距，包括一些拥有官方地位和数百万使用者的语言。然而，这些结果也表明，对参数规模在 70 亿到 90 亿之间的 LLMs 进行专门微调可以显著缩小与 GPT-4o 和 DeepSeek-R1 等大得多的最先进模型之间的差距。因此，通过关注东南亚语言的现实本地化需求，我们鼓励在数据收集、策划和该地区专用轻量级 LLM 解决方案的微调方面进行更集中的努力。

5 未来工作

我们认识到，尽管 SEA-HELM 目前覆盖菲律宾语、印度尼西亚语、泰米尔语、泰语和越南语，但我们还有更多工作要做，因此我们致力于迭代扩展每个支柱。

具体来说，我们计划扩展 SEA-HELM 以包含更广泛的语言、文化、任务和模型，以鼓励模型中更强的东南亚代表性。我们还寻求探索额外因素，如自动 LLM 评估。这将实现对更广泛东南亚语言上下文中 LLM 性能的更全面评估，尤其是该地区的低资源语言，如缅甸语、高棉语和老挝语。

局限性

虽然 SEA-HELM 旨在实现对东南亚 LLMs 的综合性和真实性评估，但在语言和任务覆盖方面尚未穷尽。我们计划未来进行迭代扩展以获得更好的覆盖。

我们的排行榜结果基于模型的单次运行。然而，由于我们假设确定性的模型行为并将每个模型的温度设置为 0，因此我们未发布结果的误差线，这与该领域的其他基准测试一致。

在 SEA-HELM 的安全支柱下，我们也承认在我们的评估中获得高分并不一定保证 LLM 在东南亚语境下的安全性，因为不可能枚举现实场景中每种不安全的回复类型。因此，通过 SEA-HELM 的安全评估必须被视为确保现实 LLM 应用安全的必要但非充分条件。

伦理声明

SEA-HELM 感谢我们的质量保证（QA）团队，他们由有偿的东南亚语言母语者组成，担任翻译和标注员，使我们能够构建本研究所需的本地化数据集。SEA-HELM 项目在经过严格的审查流程后，已获得我们大学内部审查委员会的完全正式批准，所有参与成员的薪酬和工作时间均完全符合现行大学指南和本研究所在国家的适用法规。

我们的 QA 团队通过公开广告招募，广告中明确说明了预计的工作量和时薪，与所有相关法律法规要求一致。团队主要由当地大学的学生和公众成员组成，所有人都是满足法定就业年龄要求的成年人。尽管参与者获得了参与报酬，但他们的参与完全是自愿的。任何个人可识别信息（PII）已从收集的数据中移除，不会与其身份关联。

我们没有估计他们的工作涉及接触攻击性材料的重大风险，因为他们未参与安全支柱下敏感数据的构建。尽管如此，他们被鼓励报告不当样本，并被赋予在任何时候（包括完成后）退出参与的选择，不会产生任何负面后果或利益损失。

我们不预见我们的研究会产生负面社会影响，因为我们的研究通过与母语者合作引入东南亚语言的评估数据集，充分尊重当地文化敏感性。因此，我们不认为我们的研究会助长对东南亚文化的过度概括。

致谢

本研究项目由新加坡国家研究基金会（National Research Foundation, Singapore）通过 AI Singapore 的国家大语言模型资助计划支持。

我们要向 AI Singapore 的同事们致以衷心的感谢，特别是 AI Products 和 SEA-LION 团队，感谢他们为丰富讨论做出的贡献以及慷慨分享的宝贵见解和反馈。同样，我们要感谢 HELM 团队从一开始就对我们倡议的坚定支持。

我们还要向参与本项目的东南亚语言母语者表达诚挚的感谢，感谢他们在人工评估、标注和翻译方面投入的不懈努力和时间。

参考文献

> 参考文献保留英文原文，详见原论文：https://arxiv.org/abs/2502.14301