SEA-LION：东南亚语言统一网络

翻译说明：本文翻译自论文 "SEA-LION: Southeast Asian Languages in One Network"，原文发表于 arXiv（arXiv:2504.05747）。技术术语在首次出现时保留英文并附中文解释。参考文献保留英文原文。

原文链接：https://arxiv.org/abs/2504.05747

原文作者：Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, Brandon Ong, Zhi Hao Ong, Jann Railey Montalan, Adwin Chan, Sajeban Antonyrex, Ren Lee, Esther Choa, David Ong Tat-Wee, Bing Jie Darius Liu, William Chandra Tjhi, Erik Cambria, Leslie Teo

机构：AI Singapore（新加坡人工智能研究院）、新加坡国立大学、南洋理工大学

网站：https://sea-lion.ai

摘要

近年来，大语言模型（Large Language Models, LLMs，大型语言模型）凭借其处理和生成自然语言的能力，主导了人工智能领域的大部分研究。然而，LLM 的研究和开发大多以英语为中心，导致东南亚（Southeast Asia, SEA）地区等低资源语言的代表性严重不足。为弥补这一代表性差距，我们推出了 Llama-SEA-LION-v3-8B-IT 和 Gemma-SEA-LION-v3-9B-IT 两个前沿的多语言 LLM，专为东南亚语言设计。SEA-LION 系列 LLM 支持 11 种东南亚语言，包括：英语、中

文、印尼语、越南语、马来语、泰语、缅甸语、老挝语、菲律宾语、泰米尔语和高棉语。我们的工作利用大规模多语言持续预训练（Continued Pre-Training，CPT）以及包含多阶段指令微调（Instruction Fine-Tuning, IFT）、对齐（Alignment）和模型合并（Model Merging）的综合后训练流程。在多语言基准测试中的评估结果表明，我们的模型在支持东南亚语言的 LLM 中达到了最先进的性能。我们以开源方式发布模型，以惠及更广泛的东南亚社区。

1 引言

大语言模型（LLMs）极大地推动了自然语言处理（Natural Language Processing, NLP）领域的发展，在文本生成、摘要和情感分析方面取得了卓越的性能（Brown et al., 2020; OpenAI, 2023; Dubey et al., 2024; Rivière et al., 2024; Zhang et al., 2024b; Yeo et al., 2024）。尽管 LLM 能力令人印象深刻，但其中绝大多数仍然非常以英语为中心（Wendler et al., 2024; Zhong et al., 2024）。

不幸的是，这种情况导致东南亚（SEA）等拥有众多低资源语言的地区在 LLM 方面处于劣势。菲律宾语、老挝语、缅甸语和高棉语等低资源语言未被许多以英语为中心的开源 LLM 所支持，例如 Llama（Dubey et al., 2024）和 Olmo（Groeneveld et al., 2024）。这迫切需要缩小英语与东南亚语言之间的语言资源和代表性差距。

近年来，已有许多以开源方式创建多语言 LLM 的尝试。例如，BLOOM（Scao et al., 2022）项目旨在通过支持 46 种自然语言来增加开源 LLM 中的多语言能力。Llama（Dubey et al., 2024）、Gemma（Rivière et al., 2024）和 Qwen（通义千问，Yang et al., 2024a）等主流 LLM 系列也在其最新版本中引入了多语言 LLM。在我们的评估中，我们发现这些模型在一般情况下（即基于英语数据集构建的评估基准上）的性能尚可，但在东南亚特定的基准测试上性能明显下降。

此外，研究人员也推出了 SeaLLMs（Nguyen et al., 2024; Zhang et al., 2024a）和 Sailor（Dou et al., 2024）等 LLM，专门解决东南亚语言的 LLM 差距。然而，这些模型在泰语或泰米尔语等语言上的表现仍不理想（泰米尔语是新加坡的官方语言之一，也在马来西亚等东南亚其他地区使用）（10X et al., 2024; AI Products Team, 2024）。

在本文中，我们通过提出一个具有数据透明性和可重复性的强健开源东南亚模型来解决上述问题，即 SEA-LION ——一系列在 Llama-3.1-8B-Instruct 和 Gemma-2-9B 上进行 CPT 和微调的 LLM，专注于东南亚语言。为解决性能问题，我们使用了 2000 亿个英语、代码和东南亚语言的 token（词元）以及 1680 万对英语和东南亚语言的指令-回答对，分别用于 CPT 和后训练步骤，

从而在东南亚语言上取得显著提升。为了让所有人不受限制地使用我们的模型，我们以完全开放的 MIT 许可证发布模型。

我们在 SEA-HELM（Susanto et al., 2025）和 Open LLM Leaderboard（开放 LLM 排行榜）上将我们的模型与东南亚地区类似规模的其他 LLM（如 Sailor 2（Team, 2024）和 SeaLLMs 3（Zhang et al., 2024a））进行基准测试，我们的模型达到了最先进的性能。

我们的主要贡献总结如下：

- 我们发布了两个 LLM：Llama-SEA-LION-v3-8B-IT 和 Gemma-SEA-LION-v3-9B-IT，经过精心训练以准确代表东南亚语言独特的语言多样性。
- 我们还在本文中提供了端到端训练流程的深入洞察，以惠及开发多语言 LLM 的社区。

---

2 持续预训练（CPT）

2.1 预训练数据

CPT 数据由精心策划的英语、多语言和代码语料库组成，来自多个开源数据仓库，包括 Dolma（Soldaini et al., 2024）、FineWeb（Penedo et al., 2024）、the-stack-v2（Lozhkov et al., 2024）、SEA-LION-Pile（AI Singapore, 2023）、SEA-LION-Pile-v2（AI Singapore, 2025），以及来自 CommonCrawl（CommonCrawl, 2024）和维基百科（Foundation, 2024）等公共领域的文档。

对于 SEA-LION-Pile-v2，我们使用预训练的 fastText 语言分类器（Joulin et al., 2017）从 CommonCrawl WARC 数据中筛选东南亚语言（即缅甸语、简体中文、印尼语、高棉语、老挝语、马来语、菲律宾语、泰米尔语、泰语和越南语）的文档。如果文档元数据中报告的语言代码与上述东南亚语言之一匹配，则保留该文档。此外，我们还使用 Trafilatura（Barbaresi, 2021）进一步清洗数据。

为确定 CPT 过程中东南亚语言、代码和英语之间的最佳数据集比例，我们进行了一系列小规模 CPT 实验，每次训练预算为 100 亿个 token，并使用不同比例的英语、代码和东南亚语言数据。最终，我们确定了最佳数据混合比例为：55% 东南亚语言、25% 英语和 20% 代码 token，总预算为 2000 亿个 token。各语言的详细 token 数量分布请参阅模型卡片。

2.2 CPT 流程

模型选择。 我们选择 Llama-3.1-8B-Instruct（Dubey et al., 2024）和 Gemma-2-9B（Rivière et al., 2024）作为 CPT 的基础模型。

训练设置。 参照先前工作（Dou et al., 2024），我们使用 BPE-Dropout（BPE 随机丢弃，Provilkov et al., 2020）来提高训练的性能和鲁棒性。我们使用 Warmup-Stable-Decay（WSD，预热-稳定-衰减）学习率调度器（Hu et al., 2024），预热和冷却阶段各占整个训练预算的 10%。我们使用 AdamW 优化器（Loshchilov and Hutter, 2019），最大学习率设为 1e-5，冷却后的最终学习率为 1e-7。参照 Wortsman et al.（2024），我们将 epsilon 设为 1e-15。

我们使用 Composer（Team, 2021）和 LLM Foundry（Team, 2022）进行分布式训练，采用完全分片数据并行（Fully Sharded Data Parallel, FSDP）（Zhao et al., 2023），在 Amazon Web Services（AWS）的八个 p5.48xlarge 实例节点的集群上运行。Llama 3.1 和 Gemma 2 模型的总训练时间分别约为 6 天和 10 天。

在本文中，我们将 CPT 后的模型分别称为 Llama-SEA-LION-v3-8B（Llama 3.1 持续预训练模型）和 Gemma-SEA-LION-v3-9B（Gemma 2 持续预训练模型）。

---

3 后训练

3.1 后训练数据

用于指令微调的后训练数据包括 Infinity-Instruct [Foundation 和 Chat]（Beijing Academy of Artificial Intelligence, 2024）、OpenMath-Instruct 2（Toshniwal et al., 2024）以及我们自己的 SEA-Instruct 数据集。

其中，SEA-Instruct 由多个开源指令数据集、一个按照 Magpie（Xu et al., 2024）模板合成生成的数据集以及从东南亚母语者收集的手工制作数据集组成。SEA-Instruct 和 SEA-Preference 数据集的完整详情请参阅模型卡片。

图 1：Llama-SEA-LION-v3-8B-IT 的训练流程（第 3.2.1 节）。 后训练流程包括 2 个阶段的指令微调、一个对齐阶段和多个合并阶段。虚线表示合并阶段，实线表示对齐阶段。

3.2 后训练流程

我们使用 LLaMaFactory（Zheng et al., 2024b）配合 DeepSpeed（Rasley et al., 2020）进行所有指令微调（IFT）和对齐步骤。所有 IFT 阶段均采用全模型微调（Full Model Fine-Tuning），模型来自前一步骤（第 2.2 节）和现有模型。我们使用 MergeKit（Goddard et al., 2024），所有合并步骤的权重和密度参数均设为 1。用于合并的模型根据模型许可证的开放性、合并适配性和性能进行经验性选择。

3.2.1 Llama-SEA-LION-v3-8B-IT

第 1 阶段 IFT

如图 1 所示，我们以 Llama-SEA-LION-v3-8B 为起点，使用 Infinity Instruct (Foundation)（Beijing Academy of Artificial Intelligence, 2024）和 OpenMathInstruct2（Toshniwal et al., 2024）数据集进行指令微调。两个数据集共包含约 950 万对指令对，主要为英语，围绕推理、数学和代码。我们将此阶段的模型称为 Stage-1-Llama。

第 2 阶段 IFT

我们使用 SEA-Instruct 数据集进行第二轮 IFT，该数据集包含约 730 万对指令对，其中 500 万对使用 Gemma-2-27B-Instruct（Rivière et al., 2024）模型和 Qwen2.5-32B-Instruct 模型（Yang et al., 2024a）以东南亚语言生成。其余为来自 Infinity-Instruct (Chat)（Beijing Academy of Artificial Intelligence, 2024）数据集的英语指令对。我们将此阶段的模型称为 Stage-2-Llama。

第一次合并

完成 IFT 阶段后，我们使用 DARE TIES（Yu et al., 2024; Ilharco et al., 2023）方法将 Stage-1-Llama 和 Stage-2-Llama 合并到 Llama-SEA-LION-v3-8B 中，进行第一轮合并。我们将此阶段的模型称为 Merge-1-Llama。

第二次合并

为减轻微调过程中的灾难性遗忘（Catastrophic Forgetting）（Alexandrov et al., 2024），我们进行第二轮合并，合并具有 Llama 3.1 血统的高性能指令微调模型。我们使用 Consensus TA（Wang et al., 2024b; Ilharco et al., 2023）合并方法，将原始 Llama-3.1-8B-Instruct、Llama3-8B-SEA-LION-v2.1-Instruct（SEA-LION Team, 2024）和 SuperNova-Lite（Arcee-AI, 2024）合并到 Merge-1-Llama 中。我们将此阶段的模型称为 Merge-2-Llama。

有用性和偏好对齐

我们使用 SimPO（Meng et al., 2024）和 SEA-Preference 数据集对 Merge-2-Llama 进行一轮对齐。我们将此阶段的模型称为 Aligned-SimPO-Llama。

最终合并

最后，我们使用 DELLA-Linear 合并方法。以原始 Llama-3.1-8B-Instruct 模型为合并基础，将 Merge-2-Llama 和 Aligned-SimPO-Llama 合并，产生最终模型 Llama-SEA-LION-v3-8B-IT。

3.2.2 Gemma-SEA-LION-v3-9B-IT

图 2：Gemma-SEA-LION-v3-9B-IT 的训练流程（第 3.2.2 节）。 后训练流程包括两个阶段的指令微调、一个对齐阶段和多个合并阶段。虚线表示合并阶段，实线表示对齐阶段。

第 1 阶段和第 2 阶段 IFT

与 Llama-SEA-LION-v3-8B-IT 类似，我们在 Gemma-2-9B 模型（Rivière et al., 2024）上使用相同的数据集进行两个阶段的 IFT。我们分别将两个阶段的模型称为 Stage-1-Gemma 和 Stage-2-Gemma。

第一次合并

我们使用 DELLA Linear 方法将 Gemma-2-9B-IT（Rivière et al., 2024）和 Stage-2-Gemma 合并到 Gemma-2-9B 中。我们将此阶段的模型称为 Merge-1-Gemma。

有用性和偏好对齐

以 Merge-1-Gemma 为基础模型，我们使用 SimPO 和 SEA-Preference 数据集进行一轮对齐。我们将此阶段的模型称为 Aligned-SimPO-Gemma。

最终合并

最后，以 Gemma-2-9B 模型为基础模型，我们将 Merge-1-Gemma、FuseChat Gemma-2-9B-Instruct（Yang et al., 2024b）、Gemma-SEA-LION-v3-9B 和 Aligned-SimPO-Gemma 合并，产生最终模型 Gemma-SEA-LION-v3-9B-IT。

3.3 讨论

这一后训练工作流强调在通用能力、东南亚特定语言流利性和自然对话能力之间的精心平衡。工作流中的每一步都旨在逐步优化模型，确保满足东南亚地区用户的多样化需求。

Gemma-SEA-LION-v3-9B-IT 和 Llama-SEA-LION-v3-8B-IT 的整个后训练过程分别耗时约 1350 和 1024 个 GPU 小时（在八块 H100 GPU 上）。为提高训练效率，所有后训练步骤均使用 Liger Kernel（Hsu et al., 2024），实现了约 60% 的显著内存节省。

---

4 实验设置与结果

表 1：SEA-HELM 多语言基准测试

在类似规模的指令模型上进行 NLU（自然语言理解）、NLG（自然语言生成）、NLR（自然语言推理）、NLI（自然语言推断）、指令遵循和多轮对话的评估。

模型	平均分	ID	VI	TH	TA	ID（指令遵循）	VI（指令遵循）	TH（指令遵循）	ID（MTBench）	VI（MTBench
SeaLLMs-v3-7B-Chat	39.19	42.72	48.50	42.59	12.06	57.14	53.33	47.00	59.81	65.24
Llama-3.1-8B-Instruct	41.48	51.50	51.31	45.32	15.40	77.14	75.24	63.00	56.38	57.59
Sailor2-8B-Chat	43.13	48.98	48.01	45.44	28.29	49.52	45.71	40.00	69.76	66.97
Qwen2.5-7B-Instruct	44.58	60.28	53.46	53.43	21.03	81.90	69.52	66.00	65.66	66.80
Gemma-2-9B-IT	55.33	64.04	59.86	57.22	52.28	88.57	78.10	71.00	68.78	68.37
Stage-1-Llama	50.76	51.84	51.83	46.23	27.53	69.52	73.33	59.00	42.74	46.41
Stage-2-Llama	59.49	53.87	55.18	50.92	44.80	77.14	76.19	67.00	50.90	53.72
	59.36	56.73	56.82	51.71	46.63	81.90	82.86	67.00	57.04	54.01

模型	平均分	ID	VI	TH	TA	ID（指令遵循）	VI（指令遵循）	TH（指令遵循）	ID（MTBench）	VI（MTBench
Merge-1-Llama										
Merge-2-Llama	58.01	59.19	52.63	51.89	35.40	87.62	80.95	78.00	56.38	59.32
Aligned-SimPO-Llama	51.30	54.86	51.69	46.77	26.40	82.86	80.00	68.00	68.20	64.68
Llama-SEA-LION-v3-8B-IT	**61.84**	60.50	61.48	55.92	43.61	84.76	85.71	76.00	62.65	68.32
Stage-1-Gemma	56.56	55.06	54.51	51.96	42.74	66.67	74.29	61.00	47.35	47.26
Stage-2-Gemma	66.66	64.10	61.76	56.90	57.85	89.52	82.86	76.00	60.54	58.93
Merge-1-Gemma	69.26	66.25	64.95	59.74	60.41	89.52	91.43	82.00	66.45	64.47
Aligned-SimPO-Gemma	69.37	65.69	65.47	59.51	57.38	86.67	88.57	78.00	68.89	73.67
Gemma-SEA-LION-v3-9B-IT	**69.35**	66.26	64.93	59.23	58.82	94.29	88.57	78.00	65.85	73.27

注：ID = 印尼语，VI = 越南语，TH = 泰语，TA = 泰米尔语

表 2：Open LLM Leaderboard 基准测试

在类似规模的不同指令模型上的英语性能评估（包括 IFEval、Big Bench Hard、MATH、GPQA、MuSR、MMLU-PRO 等基准）。Gemma-SEA-LION-v3-9B-IT 取得了最高的平均分 35.43。

4.1 评估设置

在评估中，我们将模型与知名 LLM 进行了比较，包括 SeaLLMs-v3（Zhang et al., 2024a）、Sailor-v2（Team, 2024）、Qwen 2.5（Yang et al., 2024a）、Gemma 2（Rivière et al., 2024）和 Llama 3.1（Dubey et al., 2024），这些模型的参数量均不超过 100 亿，与我们的模型规模相当。评估分为两个方面：

多语言性能。 我们使用 SEA-HELM 排行榜（Leong et al., 2023; Susanto et al., 2025）评估每个 LLM 的多语言性能。由于低资源语言（如老挝语、高棉语、菲律宾语）缺乏适当的基准测试，我们仅对 SEA-HELM 排行榜涵盖的语言进行了基准测试，即印尼语、泰米尔语、泰语和越南语。我们选择 SEA-HELM 是因为该基准的设计最能反映东南亚文化和知识的性能表现。我们使用了官方网站上的评估代码，未做任何修改。

英语性能。 我们使用 Open LLM Leaderboard（HuggingFace, 2024）评估模型的英语性能。该排行榜包含六个基准测试：IFEval（Zhou et al., 2023）、Big Bench Hard（Suzgun et al., 2023）、MATH（Hendrycks et al., 2021）、GPQA（Rein et al., 2023）、MuSR（Sprague et al., 2024）和 MMLU-PRO（Wang et al., 2024c）。

4.2 结果

多语言性能。 如表 1 所示，SEA-HELM 基准测试结果表明，我们的指令模型 Llama-SEA-LION-v3-8B-IT 和 Gemma-SEA-LION-v3-9B-IT 在东南亚语言上取得了有竞争力的性能，其中 Gemma-SEA-LION-v3-9B-IT 达到了最高的平均性能之一。两个模型均优于其他专注东南亚语言的 LLM，如 Sailor2-8B-Chat 和 SeaLLMs-v3-7B-Chat，在 SEA-HELM 基准测试（SEA-MTBench 任务除外）覆盖的所有语言上平均分达到 69.35。

英语性能。 Open LLM Leaderboard 的性能如表 2 所示。Llama-SEA-LION-v3-8B-IT 和 Gemma-SEA-LION-v3-9B-IT 在英语语言、数学和推理任务上均表现出色，其中 Gemma-SEA-LION-v3-9B-IT 取得了最高的平均分 35.43。

4.3 性能分析

持续预训练

CPT 阶段主要聚焦于获取东南亚语言的能力和知识。与基础模型和 CPT 模型的对比表明，Llama-SEA-LION-v3-8B 和 Gemma-SEA-LION-v3-9B 分别比 Meta-Llama-3.1-8B 和 Gemma-2-9B 在 SEA-HELM 平均性能上提升了 6.05 和 7.19。我们观察到指令遵循能力的提升

尤为显著，我们认为这是因为 CPT 模型是基于指令版模型而非基础模型进行训练的。两个 CPT 模型在 Open LLM Leaderboard 基准测试上也与 Meta-Llama-3.1-8B 和 Gemma-2-9B 基础模型表现相当，表明使用 25% 英语 token 进行训练的选择有效缓解了 CPT 带来的灾难性遗忘（Zheng et al., 2024a）。

如表 1 所示，我们选择 Gemma 是因为它在多语言基准测试上表现最好。然而，我们也证明了我们的框架对所有 LLM 均可推广——通过在 Llama 3.1 上应用我们的框架，虽然 Llama 3.1 的性能低于 Qwen 或 Sailor，我们仍能使其超越所有对手。CPT 模型和其他基础模型的完整性能分数见附录 A.1。

第 1 阶段：英语指令微调

第 1 阶段 IFT 主要聚焦于获取数学、代码和英语通用指令遵循的一般能力。虽然我们的 CPT 模型基于 Llama-3.1-8B 的指令版，但 CPT 过程已削弱了指令遵循能力（见表 2）。我们观察到 Stage-1-Llama 和 Stage-1-Gemma 分别在 IFEval 基准测试上的英语指令遵循能力提升了 3.86 和 9.72。在 SEA-HELM 基准测试上，Stage-1-Llama 和 Stage-1-Gemma 的平均提升分别为 7.9 和 7.47。

第 2 阶段：多语言指令微调

第 2 阶段 IFT 聚焦于多语言和推理能力。通过在东南亚语言和更高复杂度的英语指令对上进行指令微调，Stage-2-Llama 和 Stage-2-Gemma 在 SEA-HELM 基准测试上相比第 1 阶段模型分别平均提升了 8.73 和 10.1。

合并 1：组合第 1 和第 2 阶段

尽管第 1 和第 2 阶段取得了显著进展，但我们观察到早期阶段的灾难性遗忘效应在第 2 阶段后仍然存在。为缓解这一问题，我们将第 1 和第 2 阶段的模型合并到 CPT 模型中，之后 Merge-1-Gemma 的平均提升为 2.6。我们还观察到 Merge-1-Llama 在所有 SEA-HELM 基准测试任务上均有提升。

合并 2：融入指令模型

为了重新引入 Llama 3.1 和 Gemma 2 模型中观察到的回复有用性、相关性和信息量，我们进一步合并开源指令模型。虽然我们在越南语和泰语的 MT-Bench 基准测试分数上观察到显著提升，但也观察到 SEA-HELM 平均性能的轻微下降以及印尼语 MTBench 分数的轻微下降，我们认为这是越南语和泰语显著性能提升的可接受代价。由于 Merge-1-Gemma 已经在 SEA-HELM 基准测试上表现优异，我们选择跳过 Gemma 模型的这一步骤。

对齐步骤

在将模型与人类偏好对齐的步骤中，我们优先考虑 SEA MTBench 性能而非其他 SEA-HELM 基准测试任务。我们观察到两个模型在所有语言的 SEA MTBench 性能上均有广泛提升。然而，这伴随着指令遵循能力和印尼语整体 SEA-HELM 性能的轻微下降。

对齐步骤显著推动模型生成更长、更有帮助和更敏感的回复，但也在更多任务特定基准测试和某些语言的指令遵循方面有所折衷，我们尝试在下一步中解决这一问题。

最终合并：组合对齐模型

为补偿前几步中的能力下降，我们将 Merge-2-Llama 和 Merge-1-Gemma 与 Aligned-SimPO-Llama 和 Aligned-SimPO-Gemma 以及各自模型系列中第 3.2.1 和 3.2.2 节描述的各种开源预训练模型进行合并。

对于 Llama-SEA-LION-v3-8B-IT，我们观察到 SEA-HELM 平均性能从对齐阶段的 51.30 显著提升至 61.84，主要来自 SEA-HELM 核心任务性能的提升。这一性能提升展示了基于每个模型的优缺点进行经验性选择预训练模型进行合并以产生远超单一模型的价值。

对于 Gemma-SEA-LION-v3-9B-IT，它以更少的后训练步骤轻松达到了比 Llama-SEA-LION-v3-8B-IT 更高的性能。我们将这一性能归功于 Gemma 2 基础模型的高性能以及更大的词汇表规模（Takase et al., 2024 已证明更大的词汇表能产生更好的模型）。

---

5 结论

尽管东南亚拥有庞大的人口和丰富的语言多样性，但在开源 LLM 中仍然缺乏资源和准确的语言文化代表。在本文中，我们推出了 Llama-SEA-LION-v3-8B-IT 和 Gemma-SEA-LION-v3-9B-IT，两个基于 Llama 和 Gemma 系列 LLM、经过全面训练以在东南亚语言上达到最先进性能的多语言 LLM。SEA-LION 代表了显式支持东南亚语言的 LLM 开发的新进展。两个模型均完全开源，可用于商业用途，以增加东南亚多语言 LLM 的可及性和创新性。

附录 A

A.1 持续预训练模型的基准测试

表 3： 在类似规模的基础模型和持续预训练模型上的 SEA-HELM 多语言基准测试（NLU、NLG、NLR、NLI 和指令遵循）。

表 4： 在类似规模的不同持续预训练模型上的 Open LLM Leaderboard 基准测试。

参考文献

- 10X et al. (2024). SCB 10X, VISTEC, and SEACrowd. Thai LLM Leaderboard.
- AI Products Team (2024). AI Singapore AI Products Team. SEA-HELM.
- AI Singapore (2023). AISG AI Singapore. SEA-LION-Pile.
- AI Singapore (2025). AISG AI Singapore. SEA-LION-Pile-v2.
- Alexandrov et al. (2024). Mitigating catastrophic forgetting in language transfer via model merging. In Findings of EMNLP 2024.
- Arcee-AI (2024). Llama-3.1-SuperNova-Lite.
- Barbaresi (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In ACL 2021 System Demonstrations.
- Beijing Academy of Artificial Intelligence (2024). Infinity Instruct.
- Brown et al. (2020). Language models are few-shot learners. In NeurIPS 2020.
- CommonCrawl (2024). CommonCrawl.
- DAMO-NLP-SG (2024). SeaExam.
- Dou et al. (2024). Sailor: Open language models for South-East Asia. CoRR, abs/2404.03608.
- Dubey et al. (2024). The Llama 3 herd of models. CoRR, abs/2407.21783.
- Foundation (2024). Wikimedia Foundation. Wikimedia enterprise HTML dumps downloads.

- Goddard et al. (2024). Arcee's MergeKit: A toolkit for merging large language models. In EMNLP 2024 Industry Track.
- Groeneveld et al. (2024). OLMo: Accelerating the science of language models. In ACL 2024.
- Hendrycks et al. (2021). Measuring mathematical problem solving with the MATH dataset. In NeurIPS Datasets and Benchmarks 2021.
- Hsu et al. (2024). Liger Kernel: Efficient Triton kernels for LLM training. arXiv:2410.10989.
- Hu et al. (2024). MiniCPM: Unveiling the potential of small language models with scalable training strategies. CoRR, abs/2404.06395.
- HuggingFace (2024). Open LLM Leaderboard.
- Ilharco et al. (2023). Editing models with task arithmetic. In ICLR 2023.
- Joulin et al. (2017). Bag of tricks for efficient text classification. In EACL 2017.
- Leong et al. (2023). BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models. CoRR, abs/2309.06085.
- Loshchilov and Hutter (2019). Decoupled weight decay regularization. In ICLR 2019.
- Lovenia et al. (2024). SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In EMNLP 2024.
- Lozhkov et al. (2024). StarCoder 2 and The Stack v2: The next generation. CoRR, abs/2402.19173.
- Meng et al. (2024). SimPO: Simple preference optimization with a reference-free reward. CoRR, abs/2405.14734.
- Nguyen et al. (2024). SeaLLMs - Large language models for Southeast Asia. In ACL 2024 System Demonstrations.
- OpenAI (2023). GPT-4 technical report. CoRR, abs/2303.08774.
- Penedo et al. (2024). The FineWeb datasets: Decanting the web for the finest text data at scale. CoRR, abs/2406.17557.
- Provilkov et al. (2020). BPE-Dropout: Simple and effective subword regularization. In ACL 2020.
- Rasley et al. (2020). DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In KDD 2020.

- Rein et al. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. CoRR, abs/2311.12022.
- Rivière et al. (2024). Gemma 2: Improving open language models at a practical size. CoRR, abs/2408.00118.
- Scao et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. CoRR, abs/2211.05100.
- SEA-LION Team (2024). Llama3-8B-CPT-SEA-LIONv2.1-Instruct.
- Soldaini et al. (2024). Dolma: An open corpus of three trillion tokens for language model pretraining research. In ACL 2024.
- Sprague et al. (2024). MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In ICLR 2024.
- Susanto et al. (2025). SEA-HELM: Southeast Asian holistic evaluation of language models. arXiv:2502.14301.
- Suzgun et al. (2023). Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In Findings of ACL 2023.
- Takase et al. (2024). Large vocabulary size improves large language models. CoRR, abs/2406.16508.
- Team (2024). Sailor Team. Sailor2: Sailing in South-East Asia with inclusive multilingual LLMs.
- Team (2021). The Mosaic ML Team. Composer.
- Team (2022). The Mosaic ML Team. LLM Foundry.
- Toshniwal et al. (2024). OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. CoRR, abs/2410.01560.
- Wang et al. (2024a). SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In NAACL 2024.
- Wang et al. (2024b). Localizing task information for improved model merging and compression. In ICML 2024.
- Wang et al. (2024c). MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. CoRR, abs/2406.01574.
- Wendler et al. (2024). Do LLamas work in English? On the latent language of multilingual transformers. In ACL 2024.

· Wortsman et al. (2024). Small-scale proxies for large-scale transformer training instabilities. In ICLR 2024.

· Xu et al. (2024). Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. CoRR, abs/2406.08464.

· Yang et al. (2024a). Qwen2 technical report. CoRR, abs/2407.10671.

· Yang et al. (2024b). Weighted-reward preference optimization for implicit model fusion. CoRR, abs/2412.03187.

· Yeo et al. (2024). Self-training large language models through knowledge detection. In Findings of EMNLP 2024.

· Yu et al. (2024). Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In ICML 2024.

· Zhang et al. (2024a). SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages. CoRR, abs/2407.19672.

· Zhang et al. (2024b). Multilingual emotion recognition: Discovering the variations of lexical semantics between languages. In IJCNN 2024.

· Zhao et al. (2023). PyTorch FSDP: Experiences on scaling fully sharded data parallel. Proc. VLDB Endow., 16(12):3848–3860.

· Zheng et al. (2024a). Breaking language barriers: Cross-lingual continual pre-training at scale. In EMNLP 2024.

· Zheng et al. (2024b). LlamaFactory: Unified efficient fine-tuning of 100+ language models. In ACL 2024 System Demonstrations.

· Zhong et al. (2024). Beyond English-centric LLMs: What language do multilingual language models think in? CoRR, abs/2408.10811.

· Zhou et al. (2023). Instruction-following evaluation for large language models. CoRR, abs/2311.07911.