SEA-SafeguardBench：评估东南亚语言与文化中的AI安全

> **翻译说明**：本文翻译自论文原文 "SEA-SafeguardBench: Evaluating AI Safety in SEA Languages and Cultures"，arXiv:2512.05501。技术术语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。

作者：Panuthep Tasawong (VISTEC)、Jian Gang Ngui (AI Singapore)、Alham Fikri Aji (Google)、Trevor Cohn (Google)、Peerat Limkonchotiwat (AI Singapore)

等贡献作者；†Panuthep Tasawong 在 AI Singapore 担任访问学者期间完成本研究

摘要

Safeguard model（安全护栏模型）帮助大语言模型（LLM, Large Language Model）检测和阻止有害内容，但大多数评估仍以英语为中心，忽视了语言和文化多样性。现有的多语言安全基准测试往往依赖机器翻译的英语数据，无法捕捉低资源语言中的细微差异。

尽管东南亚（SEA, Southeast Asia）地区拥有丰富的语言多样性和独特的安全问题——从文化敏感的政治言论到区域特有的虚假信息——东南亚语言在安全研究中仍然严重不足。要弥补这些差距，需要由母语使用者原创编写的基准测试，以反映当地规范和危害场景。

我们推出了 **SEA-SafeguardBench**，这是首个经人工验证的东南亚安全基准测试，涵盖八种语言、21,640 个样本，分为三个子集：通用（general）、真实场景（in-the-wild）和内容生成（content generation）。

我们的基准测试实验结果表明，即使是最先进的LLM和安全护栏，在面对东南亚文化和危害场景时仍然表现不佳，与英语文本相比性能明显下降。

---

1 引言

大语言模型（LLM）在问答（Zhuang et al., 2023; Monteiro et al., 2024）、摘要生成（Laban et al., 2023; Li et al., 2024）和交互式对话（Zheng et al., 2023; Ameli et al., 2025）等任务上表现出色。随着LLM进入实际应用，确保其安全和负责任的行为变得至关重要。一种常见的解决方案是使用安全护栏模型来检测有害输入或过滤不安全输出，从而减少虚假信息、阻止有害行为，同时维护道德和法律标准。Han et al.（2024）表明，这种模型可以大幅防止有害回复，在英语安全基准上达到了 86.1 的 F1 分数。然而，大多数评估仍以英语为中心，这些系统能否推广到其他语言和文化环境尚不清楚，如图1A所示。

表1：基准测试对比。 提示词（prompt）和回复（response）的数量仅针对公开集提供。

现有的安全评估主要集中在英语（Vidgen et al., 2024; Röttger et al., 2024; Chao et al., 2024; Han et al., 2024; Ghosh et al., 2024, 2025; Xie et al., 2025; Cui et al., 2025; Li and Liu, 2025），只有少量数据集涉及多语言安全（Deng et al., 2024; Wang et al., 2024b; Kumar et al., 2025）。许多多语言基准测试通过机器翻译英语数据生成，验证有限。这是有问题的：机器翻译（MT, Machine Translation）系统在低资源语言上表现不佳，经常产生不准确或文化不当的翻译（Haddow et al., 2022; Merx et al., 2025; Pei et al., 2025）。因此，翻译的基准测试可能会遗漏语言和文化上的细微差别，给人以安全对齐（safety alignment）良好的误导印象。

东南亚语言在安全研究中明显不足，尽管该地区语言多样且人口超过6.71亿（占全球人口的8.75%）。目前没有原生的东南亚安全基准测试来检验那些声称支持这些语言的模型是否真正提供了安全且符合文化背景的回复。现有基准测试也以通用有害内容为中心，忽视了区域特有的问题，如文化敏感的政治言论、宗教禁忌和依赖上下文的虚假信息。东南亚安全基准测试不能简单地从英语机器翻译而来；它必须由母语使用者原创编写，以捕捉当地的危害场景、社会规范和文化敏感性。

基于以上差距，我们提出以下研究问题：

- **RQ1：语言鲁棒性**。 与英语相比，安全护栏在东南亚语言上的表现一致性如何？一个鲁棒的模型应在各语言间执行同等的安全标准。

- **RQ2：安全分类中的文化敏感性**。当前的安全护栏能否准确区分东南亚语境中文化上安全和不安全的提示，反映当地规范、禁忌和危害表达?

为解答这些研究问题，我们推出了 **SEA-SafeguardBench**，这是首个面向东南亚语境的多语言、文化细粒度安全基准测试。该基准涵盖7个东南亚国家的文化和语言：印度尼西亚（IN：印尼语）、马来西亚（MS：马来语）、缅甸（MY：缅甸语）、泰国（TH：泰语）、新加坡（TA：泰米尔语）、菲律宾（TL：塔加洛语）和越南（VI：越南语），每个实例均配有相应的英语版本。

为回答 RQ1，我们使用现有英语安全数据集中的安全和有害主题构建了**通用子集**。如图1A所示，提示词和回复通过 Google NMT（Neural Machine Translation，神经机器翻译）翻译成东南亚语言，然后由精通英语和目标语言的标注员进行编辑，所有标注员均通过了英语水平测试。

为解答 RQ2，我们在两种设置下构建了**文化子集**：（I）**真实场景（In-the-wild）**：由母语使用者编写的安全和不安全东南亚提示词，以捕捉真实世界的文化话题（图1B）。（II）**内容生成（Content generation）**：请求生成文化上不安全内容的提示词，包括虚假信息和假新闻场景，用于测试LLM能否检测和阻止此类请求（图1C）。

与之前的多语言安全基准（Deng et al., 2024; Wang et al., 2024b; Kumar et al., 2025）不同——它们通常依赖机器翻译——我们的基准测试经过完全的人工验证，确保准确性和语言忠实度。总体而言，我们的数据集包含 **13,830** 个提示词和 **7,810** 个回复，涵盖 **1,338** 个文化话题，包括本地知识、文化规范与禁忌、信仰、区域特有敏感性以及社群或群体认同。

我们在基准测试上评估了20个模型，发现当前的安全护栏模型在东南亚语言和语境上持续表现不佳，尽管在英语安全基准上表现良好。这凸显了当前模型对东南亚语境的理解和表征有限。

我们的贡献如下：

- 我们推出了 **SEA-SafeguardBench**。该基准测试包含 13,830 个提示词、7,810 个回复和 1,338 个文化话题，所有内容均经东南亚母语使用者审核。
- 与以往基准测试不同，SEA-SafeguardBench 是首个面向东南亚语境的文化基准测试，旨在研究当地规范、禁忌和危害表达。
- 我们进行了大规模实验，揭示了当前LLM和安全护栏在东南亚安全评估中的差距，发现当前模型在许多情况下仍会误分类东南亚安全话题，特别是在有害请求和有害回复方面。我们还提供了错误分析和改进分析部分，为未来研究提供方向。

---

2 SEA-SafeguardBench

图1： 我们三个子集基准测试的样例及其构建方式。我们有三个类别：(i) 全球通用安全话题，(ii) 真实场景数据集，(iii) 东南亚内容生成。

2.1 概述

表1总结了我们的 SEA-SafeguardBench 与现有基准测试之间的关键区别。大多数先前的基准测试集中于通用英语安全。多语言基准（Wang et al., 2024b; Kumar et al., 2025）大多翻译英语数据集，忽视了文化根植的风险和东南亚语言覆盖。RabakBench（Chua et al., 2025）引入了针对新加坡安全语境的文化根植提示；然而，由于其提示来源于在线论坛，该数据集主要反映人际对话而非人机交互。相比之下，SEA-SafeguardBench 直接针对东南亚文化和安全语境，提供七种东南亚语言的文化根植提示和回复，每种均配有英语用于跨语言评估。所有样本均由来自相应国家的母语使用者验证或编写，确保文化真实性和语言准确性。

2.2 通用子集

为评估当前LLM如何处理通用安全话题（RQ1），我们从三个现有基准（JailbreakBench（Chao et al., 2024）、Aegis2（Ghosh et al., 2024）和 WildGuardMix（Han et al., 2024））中各随机抽取200个实例，然后使用专业人工翻译将提示词和回复翻译成东南亚语言，如图1a所示。

> 注1：我们首先使用 Google NMT 从英语翻译到东南亚语言，以确保翻译一致性。这很重要，因为如果让所有标注员从头开始翻译而不使用 Google NMT，即使原句相同，不同标注员的翻译结果也会不同。当我们使用 Google NMT 作为初始翻译时，根据初步结果，我们发现所有标注员严格遵循指南后的最终结果几乎相同。

在我们的指南（附录A.1）中，我们让说相应东南亚语言（以及英语）的标注员编辑提示词和回复，使其更自然、正确、语法通顺。我们还允许标注员根据上下文将措辞改得更不礼貌、更具骚扰性、更自然，以接近真实场景。我们将此数据集称为**通用子集**，如表1所示。

2.3 文化集：真实场景

为评估东南亚语境中的文化理解（RQ2），仅使用翻译数据集是不够的，因为这些数据集并非设计来展示LLM是否具有对东南亚文化背景的理解。为了了解LLM在东南亚文化语境中的安全程度，我们需要一个专门设计的数据集，来衡量LLM在面对东南亚特有文化话题时，能否准确预测提示词是安全还是不安全的。

如图1b所示，我们通过提出一个专门针对AI文化相关安全评估的新子集来解决这个问题。为构建高质量且文化相关的数据，我们要求标注员写出与其国家相关的文化话题（完整的文化相关话题指南见附录A.2），最终从七个东南亚国家获得了 1,338 个话题。然后，我们要求他们基于所提供的话题，用英语和东南亚语言各写一个安全和不安全情境下的提示词。特别地，我们的标注指南允许标注员自由编写安全和不安全提示词，只要上下文与文化话题相关即可。这些提示词代表了人类在文化话题上会向AI提出的真实问题或请求。

2.4 文化集：内容生成

近年来，LLM的研究和实际应用集中于内容生成（Ayoobi et al., 2023; Acharya et al., 2023; Maleki and Zhao, 2024），包括摘要生成、博客写作和假新闻生成。大多数被测试的LLM在被提示时容易生成假新闻，包括针对东南亚文化语境的假新闻。这种不安全行为表明LLM缺乏对东南亚文化语境的充分了解，导致它们生成虚假或有害内容。因此，评估模型的这种行为有很强的必要性，因为这在东南亚地区尤其有害（RQ2）。

我们提出了一个以"不该做的事"为中心的文化内容生成数据集，使用特定的提示模板来诱导LLM在东南亚语境中创建假新闻或有害内容，如图1c所示。以下是我们构建数据集的详细方法。

提示词和回复的生成。 我们编制了每个东南亚国家"不该做的事"清单，涵盖120个话题，来源于互联网和标注员编写。然后，我们使用三个提示模板为每个条目生成提示词：(i) 提示LLM创建鼓励人们做不该做之事的内容，(ii) 提示LLM提供这些行为的操作指南，(iii) 提示LLM创建将"不该做的事"伪装成"应该做的事"的误导性内容（完整提示见附录C.1）。这为每个东南亚国家产生了360个文化根植的提示词；然后我们仅选择符合标准的提示词（即提示词和回复与话题一致，且LLM未拒绝回答）。对于每个提示词，我们使用 GPT-4o 生成英语回复。

> 注2：我们基于附录D.2的结果选择 GPT-4o，该结果显示 GPT 模型在生成东南亚自然回复方面表现最佳。

所有输出（提示词和回复）均以英语编写，然后由专业翻译人员翻译，从而实现跨语言文化理解的评估（RQ1）。

数据标注员。 虽然我们的问题基于每个国家"不该做的事"，但这并不意味着标签始终是"不安全的"，因为某些请求在东南亚国家可能是可接受的、合法的或无冲突的。为使标签与东南亚文化语境一致，四名标注员对每个提示-回复对进行标注，我们使用多数投票来确定最终标签。二元选择为：(i) 安全 和 (ii) 不安全。对于安全和不安全的标准，我们遵循先前安全护栏工作的相同方法和定义（Inan et al., 2023; Han et al., 2024），例如违反AI安全的文本，并且我们额外提出了一条新的安全规则：**文本在传统和法规方面需对该国居民具有文化适当性**（标注员指南见附录A.3）。

有趣的是，我们发现标注员在文化相关内容上的分歧比通用话题更大。例如，批评泰国王室可能被某些人认为是"安全的"，但被其他人认为是"不安全的"。

> 注3：侮辱泰国王室有法律后果，但批评王室并不违法。不过，它仍然被一些人视为不当行为，因此这是一个主观且敏感的问题。

为处理此类情况，我们引入了"**敏感**"（sensitive）标签，用于可能骚扰、冲突或冒犯群体的提示词或回复。没有明确多数意见的样本将获得此标签。标注员一致性的详细信息见附录A.4。

2.5 基准测试分析

图2： SEA-SafeguardBench 的数据统计。完整分布请参见附录A.8。

图3： 通用集和文化集的可视化。为消除语言偏差，所有样本均以英语编写，每个点代表每个国家的文化样本，而非语言。

数据统计。 图2显示了每种语言的数据统计，每个东南亚实例均配有其英语版本用于跨语言评估。数据集包含三个子集：

- (i) **通用子集**：每种语言600个提示-回复实例，共4,800个。
- (ii) **内容生成（CG）文化子集**：包含215个文化根植的英语提示-回复，由标注员翻译成东南亚语言，每种东南亚语言430个实例（215个英语 + 215个翻译），七种语言共3,010个。
- (iii) **真实场景（ITW）文化子集**：每种东南亚语言约420–480个实例，每个配有东南亚语言和英语版本（XX-EN），共6,020个。

类别分布在通用和ITW子集中是均衡的，而CG文化子集有更多敏感实例，反映了在东南亚语境中定义有害内容的挑战（标注员一致性见附录A.4）。

数据集多样性。 为检验文化样本和通用样本之间的差异，我们使用最先进的多语言模型multilingual-e5-large-instruct（Wang et al., 2024a）的嵌入向量，通过 t-SNE 绘制了所有英语样本（完整实现见附录A.7）。理想情况下，即使所有输入都是英语，文化样本也应与通用样本形成独立的聚类，反映潜在的语境差异。

图3A显示，真实场景集在文化样本和通用样本之间呈现出明显不同的聚类。我们还观察到马来-印尼语和泰国-缅甸样本的质心重叠，凸显了基准测试和真实世界语境中的区域文化亲近性。图3B显示内容生成集呈现不同模式，国家特定的聚类比真实场景集分离更明显。这是因为内容生成需要更深层的文化理解，而非依赖通用子集中的关键词线索。我们还在附录A.9中探索了句法差异。

---

3 实验设置

设置。 安全护栏评估衡量模型将输入内容分类为安全（Safe）或有害（Harmful）的能力，测量其区分可接受内容和潜在危险提示词或回复的有效性。我们在两个不同的任务上评估安全护栏：**提示词分类**和**回复分类**。由于现有安全护栏只能预测安全和有害标签，我们将敏感标签在提示词分类中映射为安全，在回复分类中映射为有害。敏感提示词被视为安全，因为它们本身并不有害，但需要谨慎处理，可以在回复生成时处理。与敏感提示词不同，敏感回复可能仍包含风险或模糊内容，因此我们保守地将其视为有害。

> 注4：为完整起见，我们在附录D.5中报告了排除敏感提示词和回复后的结果。然而，这种配置意义有限，因为处理敏感案例是确保文化安全的核心挑战。

模型。 我们评估了各种最近发布的开源和现成安全护栏的有效性，涵盖多种参数规模（模型列表见附录B）。我们还评估了最近发布的LLM的零样本（zero-shot）性能，详见附录C.2。除了安全护栏评估外，我们还报告了LLM安全评估，评估开源和API模型在有害和安全提示词上的安全回复率和拒绝率，见附录D.2。

评估指标。 与先前研究一致（Zeng et al., 2024; Inan et al., 2023），我们使用 **AUPRC**（Area Under the Precision-Recall Curve，精确率-召回率曲线下面积）评估安全护栏性能，这是一个与阈值无关的指标，评估模型在所有分类阈值范围内的性能。更高的 AUPRC 表示更有效地识别

有害输入或回复，精确率和召回率之间的权衡更好。为计算 AUPRC，我们使用代表性标记（安全和不安全）概率的置信度分数，确保跨运行的一致结果。现成的API通常返回序数类别（如低、中、高）或整数（如0–7）而非标记概率；我们将这些映射为数值以计算 AUPRC（见附录B.2）。基于阈值的指标如 F1 和 FPR（False Positive Rate，假阳性率）在附录D.5中报告。

---

4 实验结果

表2展示了20个安全护栏模型在提示词和回复分类上的性能，以回答 RQ1（跨语言鲁棒性）和RQ2（文化敏感性）。

表2：安全护栏性能（AUPRC：越高越好），用于提示词和回复分类任务。粗体值表示每个类别中的最佳模型。

模型	提示词分类			回复分类	
	通用 (EN/SEA)	ITW文化 (EN/SEA)	CG文化 (EN/SEA)	通用 (EN/SEA)	CG文化 (EN/SEA)

零样本模型：

模型	通用 EN	通用 SEA	ITW-EN	ITW-SEA	CG-EN	CG-SEA	提示词均值	通用 EN	通用 SEA	CG-EN	CG-SEA	回复均值
Gemma-3-it 4B	89.5	86.7	96.8	94.2	59.5	51.1	79.6	85.5	83.6	63.1	58.8	72.8
Gemma-3-it 27B	89.3	87.5	98.0	97.0	65.8	65.3	83.8	83.6	83.8	68.9	63.9	75.0
Gemma-SEA-LION-v4-27B	90.9	88.5	98.2	97.4	65.4	64.7	84.2	85.0	85.2	68.7	63.8	75.7
Llama-3.1-it 8B	89.8	83.8	95.1	89.4	60.3	49.9	78.1	84.1	71.3	63.2	45.5	66.0

模型	通用EN	通用SEA	ITW-EN	ITW-SEA	CG-EN	CG-SEA	提示词均值	通用EN	通用SEA	CG-EN	CG-SEA	回复均值
Llama-3.1-it 70B	90.7	87.0	97.7	94.8	67.5	62.6	83.4	87.1	83.1	65.7	59.5	73.8
Llama-3.2-it 3B	69.5	67.2	75.8	59.7	30.3	35.1	56.3	73.9	69.9	42.3	47.2	58.3
Llama-3.3-it 70B	92.0	88.1	96.8	94.3	67.9	61.2	83.4	88.3	86.3	65.9	63.0	75.9
GPT-OSS 20B	87.9	87.1	92.0	89.8	59.7	55.3	78.6	83.8	82.2	61.4	58.7	71.5
GPT-4o	**94.9**	**92.3**	**98.9**	**98.1**	65.2	59.7	**84.9**	**90.4**	**88.2**	64.5	61.7	**76.2**

微调模型：

模型	通用EN	通用SEA	ITW-EN	ITW-SEA	CG-EN	CG-SEA	提示词均值	通用EN	通用SEA	CG-EN	CG-SEA	回复均值
ShieldGemma 2B	83.1	79.9	95.8	90.6	53.2	51.8	75.7	79.1	73.3	51.5	47.3	62.8
ShieldGemma 9B	86.0	83.2	97.2	95.3	52.2	55.7	78.3	78.2	77.1	56.5	54.0	66.5
LlamaGuard-3 1B	90.1	81.6	91.8	86.4	45.7	33.9	71.6	82.8	69.5	58.6	48.6	64.9
LlamaGuard-3 8B	93.9	90.4	97.3	95.7	56.7	47.4	80.2	**92.1**	**86.8**	67.1	64.8	**77.7**
LlamaGuard-4 12B	92.6	84.6	94.6	84.7	46.0	32.4	72.5	88.1	77.2	60.9	53.6	69.9
PolyGuard-Qwen 0.5B	91.3	75.8	97.5	82.6	40.8	32.4	70.1	77.8	64.0	53.9	43.7	59.8
	92.2	85.2	98.6	94.9	53.8	41.0	77.6	80.1	77.1	67.9	61.4	71.7

模型	通用EN	通用SEA	ITW-EN	ITW-SEA	CG-EN	CG-SEA	提示词均值	通用EN	通用SEA	CG-EN	CG-SEA	回复均值
PolyGuard-Qwen 8B												
PolyGuard-Ministral 8B	93.0	88.3	98.2	95.4	53.3	42.0	78.4	87.5	81.5	67.3	61.9	74.6
Qwen3Guard-Gen 8B	**94.1**	**90.6**	96.3	95.3	55.0	45.9	79.5	91.3	89.8	**72.6**	**72.9**	**81.6**
LionGuard-2	85.6	72.7	95.8	78.5	46.7	41.9	70.2	73.9	63.5	47.8	40.3	56.4
X-Guard	84.0	80.7	97.0	86.1	42.5	35.1	70.9	-	-	-	-	-

API：

模型	通用EN	通用SEA	ITW-EN	ITW-SEA	CG-EN	CG-SEA	提示词均值	通用EN	通用SEA	CG-EN	CG-SEA	回复均值
Google Model Armor	79.1	72.5	86.6	75.6	40.1	33.8	64.6	67.2	60.7	69.4	59.1	64.1
Azure AI Content Safety	80.0	74.5	88.5	83.1	37.6	30.2	65.7	-	-	-	-	-
OpenAI Moderation	88.0	78.3	95.3	86.4	45.5	40.3	72.3	-	-	-	-	-
LakeraGuard	82.4	72.6	88.9	76.6	30.0	37.8	64.7	-	-	-	-	-

语言差距

安全护栏模型在东南亚语言上的表现始终低于英语，揭示了有限的跨语言泛化能力，特别是在类型学和语言学多样的环境中。在东南亚语言中，**泰米尔语和缅甸语**最具挑战性，在所有评估场景中都记录了最低的性能（完整结果见附录D.5）。

平均而言，所有模型的提示词分类性能在通用、ITW文化和CG文化子集上分别下降了 **5.7、6.1 和 5.4** 个 AUPRC 点。对于回复分类，我们观察到通用和CG文化子集分别下降了 **5.7 和 5.8** 个 AUPRC 点。这强调了 RQ1 中的问题，即安全护栏模型仅在某些语言上表现良好，主要是英语。定性案例示例见附录D.4。

文化差距

安全护栏模型在 ITW 文化子集上通常保持稳健的表现，该子集包含意图明确但涉及区域特有引用的提示词，如当地地标、传统节日或知名公众人物。这表明，当提示词意图明确时，仅存在区域特有实体并不会显著削弱模型性能。

然而，在 CG 文化子集上性能大幅下降，该子集需要细微的文化理解，如对当地规范、禁忌或隐性社会政治敏感性的了解。我们的评估揭示了提示词分类性能的大幅下降——英语下降 **36.4** 个 AUPRC 点，东南亚语言下降 **36.2** 个 AUPRC 点；回复分类也有类似的下降（分别为 21.0 和 21.2 个点）。这些不足揭示了当前安全护栏在理解区域特有禁忌方面的关键差距，这对于在东南亚和其他文化复杂地区的有效部署至关重要。

---

5 错误分析与改进

本节讨论如何利用现有模型的洞察来提升当前安全护栏在我们基准测试上的表现。

5.1 分类错误分析

本节我们检验：(i) 现有安全护栏的失败模式，(ii) 提供提示词作为额外上下文如何偏置回复分类。

图4展示了在我们基准测试中，对四种提示-回复对类型（{安全，有害} 提示词 × {安全，有害} 回复）进行评估的最佳安全护栏的混淆矩阵。Gemma-3-it 27B 展示出相反的过度防御模式的额外结果见图16。

失败模式。 如图4A所示，LlamaGuard-3 8B 在正常设置（有提示词访问权限）下的混淆矩阵揭示了不同的错误模式。该模型正确分类了 87% 的安全/安全（S/S）实例，在处理安全内容时表现出较强的可靠性。然而，它在有害内容上表现困难：有害/有害（H/H）实例被误分类为 S/S（25%）、S/H（4%）或 H/S（16%），41% 的 H/S 实例被误分类为 S/S。这种**防御不足**的倾向引发安全担忧，因为相当部分的不安全输入-输出被错误地接受。

一个显著的弱点出现在处理 S/H 案例中，即有害回复配有安全提示词。对于 LlamaGuard-3 8B，超过 99% 的 S/H 实例被误分类，通常误为 S/S。这表明该模型低估了从看似无害的提示词产生有害回复的风险。

提示词作为额外上下文的影响。 虽然提示词提供了上下文，但我们的基准测试使用单轮请求，用户提出问题或请求内容生成。在这些情况下，回复的有害性通常从输出本身就能看出（如明确的有害指示、虚假信息或辱骂性语言）。通过有无提示词两种条件评估回复，可以揭示安全护栏模型是依赖提示词线索还是评估生成的内容。

比较图4A和B，我们发现提示词上下文系统性地影响了回复分类： - (i) 安全提示词导致基本一致的输出，表明安全提示词不会显著偏置回复分类。 - (ii) 有害提示词增加了将回复分类为有害的可能性，无论实际安全性如何。移除提示词将 H/S→H/H 误分类从 4% 降低到 1%，但将 H/H→H/S 误分类从 16% 增加到 26%。

这些变化表明，有害提示词引入了**捷径推理**（shortcut reasoning），模型基于提示词而非仔细分析内容来标记回复为有害。

图4： 四种提示-回复对类型的混淆矩阵，分别在有（A）和无（B）提示词访问权限下进行回复分类评估。在两种设置下，提示词分类阶段都可以访问提示词。

5.2 安全护栏中阈值的最优性

安全护栏通常被框架为离散分类问题，朴素决策阈值设为 0.5（Inan et al., 2023; Zeng et al., 2024; Han et al., 2024）。在本研究中，我们认为这种常见做法可能是次优的。

图5展示了三个安全护栏模型在不同阈值下的性能。分析表明，微调的安全护栏模型（ShieldGemma 9B 和 LlamaGuard-3 8B）对阈值选择高度敏感，表现出明确的精确率-召回率权衡。F1 分数在低阈值（约0.1）处达到峰值，随着阈值增加而恶化。这一发现表明，使用固定 0.5 阈值的常见做法通常是次优的，可能会显著低估模型性能。

相比之下，零样本安全护栏模型 Gemma-3-it 27B 对阈值变化的敏感性极低，倾向于**召回率优先于精确率**。这种召回率导向的行为限制了可调性，通常导致过度标记输入为不安全，减少了有害内容，但牺牲了实际使用效果。

图5： 安全护栏在不同阈值下的提示词分类（上）和回复分类（下）性能。

5.3 模型在模糊案例上的行为

SEA-SafeguardBench 将提示词和回复分为三种类型：安全、敏感和有害。敏感类别代表既非明确安全也非明确有害的模糊案例。我们检验三个安全护栏模型如何对这种模糊性进行评分，期望它们分配中等置信度，而非将敏感内容视为明确安全或有害。

图6揭示，没有一个模型在处理敏感提示词和回复时表现出这种不确定性。它们不是分配中等置信度分数，而是频繁产生**过度自信的预测**，将敏感内容视为明确安全或明确有害。这一发现凸显了当前安全护栏模型的一个关键局限：它们在面对模糊内容时无法表达校准的不确定性。这种行为在需要细致安全判断的实际场景中可能导致误分类并降低可信度。

图6： 不同提示类型的提示词（上）和回复（下）分类置信度分数分布。

5.4 文化感知安全护栏

我们研究了将文化感知融入模型以提高文化敏感样本性能的影响。我们在三个零样本安全护栏模型——Gemma-SEA-LION-v4-27B、Llama-3.3-it 70B 和 GPT-4o——上进行实验，通过添加指令要求模型在分类时考虑相应国家的文化规范（完整实现细节见附录D.6）。

如表3所示，对于已经熟悉目标文化的模型（如 Gemma-SEA-LION-v4-27B），融入文化感知带来了明显的性能提升。相比之下，没有接触过该文化背景的模型（即 Llama-3.3-it 和 GPT-4o）仅表现出微小或不一致的改善，表明**仅靠文化指令而没有底层的区域特定预训练知识是不够的**。然而，当模型在文化相关数据上进行过预训练时——即 SEA-LION，其包含大量东南亚预训练文本——尽管从未在安全数据上训练过，它在文化安全基准上也取得了显著提升。

表3： 添加文化感知提示后在CG子集上的性能变化（对比表2）。

6 相关工作

6.1 安全基准测试

现有的LLM安全基准测试以英语为主，目标行为包括有害内容审核（如 OpenAIModeration（Markov et al., 2023）、SimpleSafetyTests（Vidgen et al., 2024）、ToxicChat（Lin et al., 2023）、BeaverTails（Ji et al., 2023））、过度拒绝（如 SORRY-Bench（Xie et al., 2025）、OR-Bench（Cui et al., 2025）、XSTest（Röttger et al., 2024））和越狱鲁棒性（如 JailbreakBench（Chao et al., 2024））。少数基准如 WildGuardMix（Han et al., 2024）旨在更广泛地覆盖。

多语言基准已经开始出现（如 XSafety（Wang et al., 2024b）、PolyGuard（Kumar et al., 2025）、MultiJail（Deng et al., 2024）、SEALBench（Shan et al., 2025）），但它们主要依赖翻译的英语数据集，缺乏文化根植的不安全内容。近期工作纳入了本地化数据（Chua et al., 2025; Ng et al., 2024），但仍然有限，集中于仇恨言论而非通用LLM安全。尽管有这些进展，仍然需要一个超越表面多语言性、捕捉多样文化规范和敏感性的基准测试。

6.2 LLM中的安全性

实现LLM安全的常见技术是执行 SFT（Supervised Fine-Tuning，有监督微调）后接 RLHF（Reinforcement Learning from Human Feedback，基于人类反馈的强化学习）（Ouyang et al., 2022; Glaese et al., 2022; Bai et al., 2022），但这种方法需要高昂的人工监督成本。近期工作（Song et al., 2025; Zhao et al., 2025）探索使用奖励信号进行多语言安全对齐，但评估仍限于翻译或高资源数据集。

另一方面，研究人员提出了在推理时过滤不安全内容的安全护栏模型，通常作为模块化安全层运行；然而，大多数现有模型仅在英语上训练和评估（Inan et al., 2023; Zeng et al., 2024; Ghosh et al., 2024, 2025; Han et al., 2024）。PolyGuard（Kumar et al., 2025）通过结合翻译和真实场景样本的17种语言数据集扩展了覆盖范围，近期工作使用翻译的英语数据集针对东南亚语言（Tan et al., 2025; Shan et al., 2025）。尽管取得了进展，大多数多语言安全护栏模型仍依赖机器翻译数据，无法捕捉文化特定的危害表达。

7 结论

我们推出了 **SEA-SafeguardBench**，这是首个面向东南亚的文化根植多语言安全基准测试。与以往主要关注语言理解的工作不同，我们的基准测试在安全关键场景中同时评估语言能力和文化能力。

我们的实验表明： 1. 模型在文化细微的安全风险方面仍然力不从心； 2. 它们往往无法将敏感内容与明确安全或有害的内容区分开来； 3. 将安全护栏视为固定阈值分类任务会导致次优结果； 4. 提高安全性、实用性和文化理解需要同时增强安全护栏模型和对齐的LLM。

这些发现揭示了当前安全方法的关键局限。我们希望 SEA-SafeguardBench 能够推动更具文化包容性的安全研究，支持AI在代表性不足的地区负责任地部署。

致谢

本研究得到新加坡国家研究基金会"国家大语言模型资助计划"的支持。本文中表达的任何观点、发现和结论或建议均为作者观点，不代表新加坡国家研究基金会的立场。

局限性

与其他低资源数据收集项目（Lovenia et al., 2024; Winata et al., 2025; Ng et al., 2025; Cahyawijaya et al., 2025）类似，我们的工作也集中于东南亚地区的主要语言和国家，包括泰国、越南、菲律宾、缅甸、新加坡、印度尼西亚和马来西亚。我们承认未纳入的其他国家包括文莱、老挝、柬埔寨和东帝汶。我们无法找到通过指南资格认定的标注员来标注这些语言的基准测试。然而，我们强调我们的基准测试可以扩展到这些语言，因为我们已经在非拉丁文字语言（如泰语和缅甸语）上开展了工作。如果标注员可用，扩展到老挝语和高棉语是可能的。

与其他基准测试工作类似（Lovenia et al., 2024; Winata et al., 2025; Ng et al., 2025; Cahyawijaya et al., 2025; Deng et al., 2024; Wang et al., 2024b），我们没有提出缓解东南亚安全问题的新模型。然而，我们在整个第5节中讨论了如何在我们的基准测试上取得高分。我们提供了分类错误和文化敏感性研究，为未来在东南亚安全问题上的研究提供了有趣的方向。

伦理声明

关于标注员详情，如附录A.5所述，我们雇用了50名说东南亚语言的标注员。我们随后进行了标注实验，仅选择通过标注测试的标注员。此外，每位标注员的报酬为 18 美元/小时，高于平均水平。我们还要求标注员在标注前考虑数据的敏感性，因为我们数据集中的某些样本可能对他们过于敏感。标注员可以在不舒适时自由退出。

参考文献

- Acharya et al. (2023). Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. Llm based generation of item-description for recommendation system. In Proceedings of the 17th ACM conference on recommender systems, pages 1204–1207.
- Achiam et al. (2023). Josh Achiam et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Ameli et al. (2025). Siavash Ameli et al. A statistical framework for ranking LLM-based chatbots. In ICLR 2025.
- Ayoobi et al. (2023). Navid Ayoobi et al. The looming threat of fake and llm-generated linkedin profiles. In Proceedings of the 34th ACM conference on hypertext and social media, pages 1–10.
- Azure (2025). Azure ai content safety documentation.
- Bai et al. (2022). Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint, arXiv:2204.05862.
- Cahyawijaya et al. (2025). Samuel Cahyawijaya et al. Crowdsource, crawl, or generate? creating SEA-VL. In ACL 2025, pages 18685–18717.
- Chao et al. (2024). Patrick Chao et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. Preprint, arXiv:2404.01318.
- Chua et al. (2025). Gabriel Chua et al. Rabakbench: Scaling human annotations to construct localized multilingual safety benchmarks. Preprint, arXiv:2507.05980.

- Cui et al. (2025). Justin Cui et al. Or-bench: An over-refusal benchmark for large language models. Preprint, arXiv:2405.20947.
- Deng et al. (2024). Yue Deng et al. Multilingual jailbreak challenges in large language models. Preprint, arXiv:2310.06474.
- Gemma Team (2024). Google Gemma Team. Gemma.
- Gemma Team (2025). Google Gemma Team. Gemma 3.
- Ghosh et al. (2024). Shaona Ghosh et al. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. Preprint, arXiv:2404.05993.
- Ghosh et al. (2025). Shaona Ghosh et al. Aegis2.0: A diverse ai safety dataset and risks taxonomy. Preprint, arXiv:2501.09004.
- Glaese et al. (2022). Amelia Glaese et al. Improving alignment of dialogue agents via targeted human judgements. Preprint, arXiv:2209.14375.
- Google Cloud (2025). Model armor overview.
- Haddow et al. (2022). Barry Haddow et al. Survey of low-resource machine translation. Computational Linguistics, 48(3):673–732.
- Han et al. (2024). Seungju Han et al. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In NeurIPS 2024, volume 37, pages 8093–8131.
- Inan et al. (2023). Hakan Inan et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv preprint arXiv:2312.06674.
- Ji et al. (2023). Jiaming Ji et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. Preprint, arXiv:2307.04657.
- Kumar et al. (2025). Priyanshu Kumar et al. Polyguard: A multilingual safety moderation tool for 17 languages. Preprint, arXiv:2504.04377.
- Laban et al. (2023). Philippe Laban et al. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In EMNLP 2023, pages 9662–9676.
- LakeraAI (2025). Lakeraguard.
- Li et al. (2024). Dongyuan Li et al. Active learning for abstractive text summarization via LLM-determined curriculum. In Findings of EMNLP 2024, pages 8959–8971.

- Li and Liu (2025). Hao Li and Xiaogeng Liu. Injecguard: Benchmarking and mitigating over-defense in prompt injection guardrail models. Preprint, arXiv:2410.22770.
- Lin et al. (2023). Zi Lin et al. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. Preprint, arXiv:2310.17389.
- Llama Team (2024). AI @ Meta Llama Team. The llama 3 herd of models. Preprint, arXiv:2407.21783.
- Lovenia et al. (2024). Holy Lovenia et al. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In EMNLP 2024, pages 5155–5203.
- Maleki and Zhao (2024). Mahdi Farrokhi Maleki and Richard Zhao. Procedural content generation in games: A survey with insights on emerging llm integration. In AIIDE 2024, volume 20, pages 167–178.
- Markov et al. (2023). Todor Markov et al. A holistic approach to undesired content detection in the real world. Preprint, arXiv:2208.03274.
- Merx et al. (2025). Raphael Merx et al. Low-resource machine translation: what for? who for? In LoResMT 2025, pages 54–65.
- Monteiro et al. (2024). João Monteiro et al. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. In NeurIPS 2024.
- Ng et al. (2025). Raymond Ng et al. Sea-lion: Southeast asian languages in one network. Preprint, arXiv:2504.05747.
- Ng et al. (2024). Ri Chi Ng et al. SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore. In WOAH 2024, pages 312–327.
- OpenAI (2024). Upgrading the moderation api with our new multi-modal moderation model.
- OpenAI (2025). Introducing gpt-oss.
- Ouyang et al. (2022). Long Ouyang et al. Training language models to follow instructions with human feedback. Preprint, arXiv:2203.02155.
- Pei et al. (2025). Renhao Pei et al. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In ACL 2025, pages 8767–8788.

- Röttger et al. (2024). Paul Röttger et al. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In NAACL 2024, pages 5377–5400.
- Shan et al. (2025). Wenliang Shan et al. Sealguard: Safeguarding the multilingual conversations in southeast asian languages. Preprint, arXiv:2507.08898.
- Song et al. (2025). Jiayang Song et al. Multilingual blending: Large language model safety alignment evaluation with language mixture. In Findings of NAACL 2025, pages 3433–3449.
- Tan et al. (2025). Leanne Tan et al. Lionguard 2: Building lightweight, data-efficient & localised multilingual content moderators. Preprint, arXiv:2507.15339.
- Team et al. (2023). Gemini Team et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.
- Upadhayay et al. (2025). Bibek Upadhayay et al. X-guard: Multilingual guard agent for content moderation. Preprint, arXiv:2504.08848.
- Vidgen et al. (2024). Bertie Vidgen et al. Simplesafetytests: a test suite for identifying critical safety risks in large language models. Preprint, arXiv:2311.08370.
- Wang et al. (2024a). Liang Wang et al. Multilingual e5 text embeddings: A technical report. Preprint, arXiv:2402.05672.
- Wang et al. (2024b). Wenxuan Wang et al. All languages matter: On the multilingual safety of large language models. Preprint, arXiv:2310.00905.
- Wang et al. (2025). Xunguang Wang et al. Sok: Evaluating jailbreak guardrails for large language models.
- Winata et al. (2025). Genta Indra Winata et al. (cited in Limitations section).
- Xie et al. (2025). (SORRY-Bench, cited in Related Works).
- Zeng et al. (2024). (cited in Metrics section).
- Zhao et al. (2025). (cited in Related Works).
- Zheng et al. (2023). (cited in Introduction).
- Zhuang et al. (2023). (cited in Introduction).