SEA-Guard：基于文化背景的东南亚多语言安全防护模型

> **翻译说明**：本文翻译自论文 "SEA-Guard: Culturally Grounded Multilingual Safeguard for Southeast Asia"，原文发表于 arXiv:2602.01618。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)

摘要

具有文化意识的安全防护（safeguard）对于 AI 在真实场景中的对齐（alignment，即确保AI行为符合人类价值观）至关重要。在这些场景中，安全不仅包括常识，还涵盖多样化的本地价值观、社会规范和地区性法规。然而，由于资源有限和母语标注者稀缺，构建大规模、基于文化背景的数据集极具挑战性。因此，许多安全防护模型依赖于对英文数据集的机器翻译，往往忽略了地区和文化方面的细微差异。

我们提出了一种新颖的智能体（agentic）数据生成框架，可以可扩展地创建真实的、针对特定地区的东南亚（Southeast Asia，SEA）安全数据集。在此基础上，我们推出了 SEA-Guard 系列——首个基于东南亚文化背景的多语言安全防护模型。在多个基准测试和文化变体的评估中，SEA-Guard 在检测地区敏感或有害内容方面始终优于现有安全防护模型，同时保持了强大的通用安全性能。

1 引言

安全防护模型（safeguard model）被部署在大语言模型（Large Language Model, LLM）的前端或后端，用于将提示（prompt）和回复（response）分类为安全或有害。通过安全防护模型，我们可以阻止用户提交敏感或不安全的提示，并阻止 LLM 返回不安全的输出（图1）。

此前的工作（Inan et al., 2023; Zeng et al., 2024; Shan et al., 2025）已在 LLM 部署系统中实施了安全防护，提高了用户安全性。实验结果也显示出强大的安全性能，特别是在英语安全基准测试上（Han et al., 2024; Chao et al., 2024a），而多语言安全——尤其是低资源语言的安全——仍未被充分探索。

> **图1**：安全防护模型的部署位置及其保护 LLM 的方式示意图。

大多数现有安全防护主要为英语设计（Inan et al., 2023; Zeng et al., 2024），很少涉及多语言场景（Kumar et al., 2025; Shan et al., 2025; Tan et al., 2025）。这些多语言安全防护通常使用在翻译数据集上训练的大型 LLM（Upadhayay and Behzadan, 2025; Kumar et al., 2025; Verma et al., 2025; Shan et al., 2025）。然而，机器翻译对许多东南亚语言表现不佳，且常常忽略文化敏感的东南亚话题（如食物、传统、历史和地方特色），导致在这些内容上的表现较弱。考虑到东南亚约占全球人口的10%，这一局限性尤其令人担忧。

为了揭示当前安全防护中的文化理解缺口，我们展示了一个文化理解在现实场景中至关重要的例子。如图1中的文化示例所示，一个假设所有印度尼西亚人都是穆斯林的提示未被当前最先进（State-of-the-Art, SOTA）安全防护（Zeng et al., 2024）拦截，导致 LLM 向用户返回了有害回复。此类情况需要基于文化的知识和多语言支持——这些能力即使在 SOTA 安全防护中仍然缺失。

基于上述考虑，我们提出三个研究问题，以系统分析现有安全防护模型的局限性，并指导为东南亚语言和文化开发稳健的安全防护：

- **RQ1：多语言一致性**。安全防护在不同东南亚语言之间能在多大程度上实现一致的安全性能?
- **RQ2：基于文化的知识**。安全防护在处理文化敏感话题时，能在多大程度上捕捉和应用东南亚文化知识?
- **RQ3：对未知领域的泛化能力**。安全防护对训练中未见过的领域泛化能力如何?

为解决上述研究问题，我们提出了 SEA-Guard——一个在8种东南亚语言的文化数据上训练的东南亚安全防护模型：缅甸语、英语、他加禄语、印尼语、马来语、泰米尔语、泰语和越南语，代

表东南亚8个国家。SEA-Guard 使用一种新颖的东南亚特定数据合成框架构建，该框架通过多个智能体和 LLM 生成文化安全数据集。

我们的合成框架通过两个新颖组件区别于其他工作：（i）文化安全数据生成——所有样本都是与东南亚话题相关的文化细微差异样本；（ii）智能体数据标注流程——用于标注和验证，过滤低质量、无效模式和重复样本。最终数据集每种语言包含 87万个样本，涵盖53个东南亚文化类别（如食物、节日、传统、政治）。利用这个精选数据集，我们训练了三个模型变体：SEA-Guard-4B、-8B 和 -12B。

为评估 SEA-Guard，我们在与研究问题对应的三个基准上进行实验：（i）东南亚安全基准（对应 RQ1 和 RQ2），（ii）通用多语言安全基准（对应 RQ1 和 RQ3），（iii）使用视觉-文本安全基准的零样本（zero-shot）任务和领域（对应 RQ3）。

结果显示，SEA-Guard 在文化安全基准上达到了最先进水平，在通用安全方面也保持了竞争力——尽管未在通用安全防护数据上训练。SEA-Guard 还能泛化到未知的视觉-语言基准，在7个案例中的6个上提升了基线。进一步分析表明，SEA-Guard 对防御不足和过度防御问题以及对抗攻击具有鲁棒性。我们将在 CC-BY-SA 许可下发布所有成果。

本工作的贡献如下：

- 我们提出了 SEA-Guard——专为东南亚地区设计的 SOTA 安全防护，提供三种大小：4B、8B 和 12B。
- 我们提出了一个数据合成框架，用于生成东南亚文化提示、回复和安全标签。最终结果为每种东南亚语言87万个样本。
- 我们采用大规模评估来回答 RQ1-3，使用了多种文本和视觉-文本数据集，包括三项分析研究。

2 SEA-Guard

> **图2**：SEA-Guard 文化训练数据构建流程示意图。我们将数据生成框架分为四个部分，各部分的详细信息在相应章节中说明。

2.1 概述

要构建一个适用于东南亚场景的稳健安全防护模型，该模型必须在东南亚特定的文化知识上进行训练。由于东南亚文化和语言数据集的不可用，我们需要构建东南亚文化安全数据集。

此前的数据合成框架（Yang et al., 2024; Deng et al., 2025; Joshi et al., 2025）表明 LLM 可以生成高质量的训练数据。与这些工作不同的是，我们的目标是一个文化多样、多语言、以安全为重点的数据集，这要求 LLM 在低资源语言中生成和标注（安全或有害）内容。因此，我们需要设计一个与研究问题（RQ1-3）对齐的新数据合成框架。

如图2所示，SEA-Guard 通过数据和模型构建中的5个主要组件区别于先前工作：

- **输入构建**（2.2节）：描述如何创建需求和指南，以引导 LLM 生成我们需要的文化样本。
- **提示和回复生成**（2.3节）：说明如何将指南、人物角色和目标语言整合到 LLM 中，以生成东南亚文化提示和回复。
- **数据标注与质量保证**（2.4节）：描述用于标注生成数据和自动确保数据质量的方法。
- **SEA-Guard 训练**（2.5节）：讨论模型选择和训练以构建 SEA-Guard-4B、-8B 和 -12B。

2.2 输入构建

与先前工作（Yang et al., 2024; Deng et al., 2025; Joshi et al., 2025）不同，我们的数据合成框架超越了直接提示，通过明确指定目标和生成指南，确保覆盖东南亚地区的语言（RQ1）和文化（RQ2）方面。

如图2A所示，我们使用与东南亚场景相关的四个元数据维度来定义需求：（i）文化主题，（ii）国家，（iii）提示类型，（iv）标签类型。我们优先处理样本较少的元数据组合，以保持数据集平衡。

指南智能体（guideline agent）根据指定的主题和需求生成逐步指南。这些指南参照人工标注协议，包括：（i）主题和目标，（ii）任务分解类别（如敏感度级别），（iii）数据规范（如元数据），（iv）示例，（v）安全伦理（如禁止行为），（vi）说明，（vii）验证。通过这种精细的指导，我们可以仔细地构建与目标对齐的提示。

2.3 提示和回复生成

为生成提示和回复，我们使用上一步获得的指南，并结合人物角色（persona）和目标语言。具体来说，我们添加了人物角色（即生活在特定国家的特定年龄和性别的人）和目标语言（因为东南亚一些国家使用多种语言）。

这是因为文化安全数据集比一般合成数据集需要更多信息，特别是在共享文化和规范的地区。例如，泼水节（Songkran）在泰国和缅甸有所不同：缅甸的佛教沐浴仪式在泼水节开始时举行，而泰国在结束时举行——前者在泰国的语境下是不恰当的。因此，结合指南、人物角色和语言有助于 LLM 更准确地捕捉东南亚特定的文化背景。

如图2B所示，我们使用 Gemma-SEA-LION-v4-27B（Ng et al., 2025）构建了提示生成器智能体，利用包含指南、人物角色和目标语言的系统和指令提示来生成英语和东南亚语言的提示。在每次生成时，我们通过改写提示进行数据增强（data augmentation），以减轻关键词偏差（keyword bias）（Ren and Xiong, 2023; Tasawong et al., 2025a），因为同一主题的提示往往具有相似的模式。

对于回复生成，我们使用四个 LLM（Llama3.1-70B-IT、Gemma3-27B-IT、Gemma-SEA-LION-v4-27B-IT 和 GPT-OSS-20B-IT）来生成多样化的回复。

2.4 数据标注与质量保证

在精心构建文化提示及其回复之后，我们需要对每个生成的样本进行标注和质量评估。为此，我们采用了蒙特卡洛推理集成（Monte Carlo Reasoning Ensemble, MCRE）技术（2.4.1节），该技术适用于数据标注（2.4.2节）和验证（2.4.3节），如图2C所示。

2.4.1 蒙特卡洛推理集成（MCRE）

为大规模文化细微差异安全分类的训练数据进行标注和验证面临三个挑战：（i）可扩展性——数据量使人工标注不可行；（ii）标注准确性——用于可靠监督；（iii）不确定性建模——即为模糊或边界案例分配软标签或概率标签。

常见解决方案是使用链式思维（Chain-of-Thought, CoT）LLM 进行零样本标注（Tan et al., 2025; Wei et al., 2022）。然而，先前关于文化安全的工作（Tasawong et al., 2025b）表明，此类模型往往过度自信，且来自单一推理轨迹的概率无法很好地捕捉真正的不确定性，限制了其处理边界和文化细微差异案例的能力。

为解决这些挑战，我们提出了用于稳健零样本分类的蒙特卡洛推理集成（MCRE），该方法对每个输入执行多次随机推理，以探索多样化的推理轨迹，并将结果预测聚合为最终分类。

对于每个输入实例 x，我们执行 N 次独立的随机推理以获得推理轨迹集合：

$R = \{r_1, \ldots, r_n\}, r_i \sim P(r|x)$ (1)

设 C 表示候选类别集合。每条推理轨迹 r_i 产生一个预测类别 $\hat{y}_i \in C$，从条件分布 $P(\hat{y}|r_i, x)$ 中采样。这些预测共同形成一个集成 $\{\hat{y}_1, \ldots, \hat{y}_n\}$，捕捉模型在随机推理过程中的预测变异性。

对于每个类别 $c \in C$，最终的类别概率估计为 c 在集成中的经验频率：

$P(\hat{y}\_final = c \mid R, x) = (1/N) \Sigma_{i=1}^{N} I(\hat{y}_i = c), c \in C$ (2)

这种聚合产生了 C 上的归一化类别概率分布，明确捕捉了随机推理引起的预测不确定性。我们可以将此技术用于每个实例 x 的标注和验证。

2.4.2 提示和回复标注

对于每对提示-回复对，我们使用 MCRE 方法（N=10）标注：（i）提示安全标签，（ii）回复安全标签，采用三类安全分类：安全（Safe）、敏感（Sensitive）和有害（Harmful）。

这里 x 表示被标注的输入实例：对于提示标注，x 对应单独的提示；对于回复标注，x 对应完整的提示-回复对。

我们不直接使用 MCRE 预测三类安全标签，而是在五类有序空间上进行分类：C_safety = {安全，安全-敏感，敏感，敏感-有害，有害}。这种设计提供了一个中间标注空间，允许模型在边界案例中表达不确定性——这些案例中安全与敏感、或敏感与有害之间的区分本质上是模糊的。

为将预测的五类有序分布映射回目标三类分类，我们首先计算连续的有害性分数 $h(x)$。具体来说，我们为每个有序标签分配归一化的严重性分数 $s_n \in [0,1]$，值均匀分布以反映递增的有害性：安全(0.0)、安全-敏感(0.25)、敏感(0.5)、敏感-有害(0.75)、有害(1.0)。有害性分数定义为预测分布下的期望严重性：

$h(x) = \Sigma\_\{c \in C\_safety\} s_v \cdot P(\hat{y}\_final = c \mid R, x)$ (3)

最后，我们使用固定阈值将连续的有害性分数离散化为三级安全标签：

- Label(x) = 安全，当 $h(x) < 0.33$
- Label(x) = 敏感，当 $0.33 \leqslant h(x) \leqslant 0.66$
- Label(x) = 有害，当 $h(x) > 0.66$

虽然该方法对于文化细微差异的安全评估很有效，但每个输入需要 N 次随机推理生成会产生大量开销——比单次反思式安全防护慢两个数量级以上——使得该方法不适合实时使用。这一成本在离线场景中是可接受的，该方法非常适合标注大规模数据集。

2.4.3 数据质量保证

为验证生成的提示是否满足指定的要求，我们沿四个维度评估每个提示：（i）要求的安全级别与标注的安全级别之间的对齐；（ii）与指定文化背景的一致性；（iii）主题相关性；（iv）与预期用途的一致性。

我们使用三个额外的零样本分类器——文化分类器、主题分类器和用途分类器——每个都使用 MCRE 方法（N=10）实现。

我们过滤掉以下样本：（i）要求的安全标签与标注的安全标签不匹配；（ii）违反指定的文化背景；（iii）同时不匹配指定的主题和预期用途。仅在主题或用途之一不匹配的样本被保留，因为在灵活解释需求的情况下它们可能仍然有效。此过程产生每种东南亚语言100万个过滤后的样本。

2.4.4 数据去重

先前工作（Tasawong et al., 2025a）表明，合成安全数据集通常包含具有重复结构的近重复样本；例如，安全示例经常被表述为问题，而有害示例则以祈使句命令出现。这种重复引入了虚假相关性（spurious correlations）（Wang et al., 2022; Hughes et al., 2024; Ye et al., 2025），并在不增加语义多样性的情况下膨胀了数据集大小。

为解决这一问题，我们识别并移除可以被简单的词袋（bag-of-words）分类器自信预测的无信息训练样本。我们采用词袋模型是因为它捕捉表面的词汇线索而有意忽略语义结构，使其非常适合检测捷径模式。此类样本可能编码了虚假相关性，移除它们可以减少训练数据中的冗余模式，同时不改变整体标签分布。

使用此程序，我们将数据集从每种东南亚语言100万个样本精简到87万个，在保持数据集覆盖率的同时减轻了重复模式。

2.4.5 人工验证

最后，为验证训练数据质量，我们雇用了32名在各自东南亚国家长大的母语标注者来验证提示和回复质量，每位标注者审查100个样本。

我们发现 79.51% 的样本质量较高，标签正确、内容准确、写作自然且语法正确。另有 12.25% 的样本在写作质量上处于边界，但安全标签正确。仅有 8.24% 的样本在写作和标签正确性方面质

量较低（大多数低质量样本来自缅甸语，其中泰语、英语和缅甸语之间偶尔的语码转换导致了错误标注）。

我们强调，由于这是合成训练数据集而非测试数据，标签正确性比写作质量更为重要。

2.5 SEA-Guard 训练

为构建适用于东南亚场景的稳健安全防护，我们选择了为该地区训练和优化的基础模型。遵循先前工作（Shan et al., 2025; Kumar et al., 2025; Zhao et al., 2025），我们根据 SEA-HELM（Susanto et al., 2025）——一个评估东南亚语言和文化理解能力的基准——选择在东南亚语言上表现良好的模型。

Qwen-SEA-LION-v4-VL（4B 和 8B）和 Gemma3-12B 在东南亚文化和对话基准上都取得了强劲的性能。相应地，我们将它们作为基础模型：SEA-Guard-4B、SEA-Guard-8B 和 SEA-Guard-12B。（我们也在10万个样本上训练了其他模型如 Gemma3-4B、Llama-3 和 Llama-SEA-LION，但只有所选模型在测试集上表现良好。）

虽然现有安全防护（如 Qwen3Guard、ShieldGemma）可以作为基础模型，但它们底层的安全策略不透明，可能引入未知偏差。

3 实验设置

对比方法。我们将模型与相同或相似大小的现有安全防护进行比较。我们评估了多个版本的 ShieldGemma（Zeng et al., 2024）、LlamaGuard（Inan et al., 2023）、PolyGuard（Kumar et al., 2025）、LionGuard-2（Tan et al., 2025）、X-Guard（Upadhayay et al., 2025）和 Qwen3Guard（Zhao et al., 2025）。这些模型基于在安全数据集上微调的 LLM（如 Llama3、Gemma2、Qwen3）。

我们还评估了安全防护 API，如 Google Model Armor（Google Cloud, 2025）、Azure AI Content Safety（Azure, 2025）、OpenAI Moderation（OpenAI, 2024）和 LakeraGuard（LakeraAI, 2025）。

基准测试和指标。我们使用为东南亚场景设计或适用的安全基准评估模型：

- **SEA-SafeguardBench**（Tasawong et al., 2025b）：一个通用但文化敏感的基准（包含 In-the-Wild 和 Content Generation 子集），专为东南亚文化开发。

- **SEALS**（Shan et al., 2025）：从 WildGuardMix（Han et al., 2024）使用 Google Translate 翻译的通用安全基准，未经人工验证。
- **SafeQA**（Ji et al., 2025）：一个通用回复安全基准，每个实例使用人机联合标注。
- 视觉-文本安全基准：**VSCBench**（Geng et al., 2025）、**VLGuard**（Zong et al., 2024）、**MSSBench-Chat** 和 **MSSBench-Embodied**（Zhou et al., 2025）。

遵循先前工作（Inan et al., 2023; Zeng et al., 2024），我们使用 AUPRC（精确率-召回率曲线下面积）作为所有基准的主要指标。

4 实验结果

4.1 东南亚文化安全结果

如表1所示，SEA-Guard-12B 在提示和回复分类上均取得了最佳性能，分别为 79.5 和 75.2。虽然 SOTA 基线 ShieldGemma 在提示分类上达到了 75.1，但在回复分类上表现明显更差（55.2），两项任务之间差距达 19.9 分。相比之下，SEA-Guard 表现出更小的差距，表明更高的可靠性和泛化能力。

SEA-Guard-4B 在提示分类上也优于同等规模的4B和8B模型，在回复分类上与 Qwen3Guard-Gen 8B 仅有 0.1 分的差距。在所有东南亚语言中，SEA-Guard 的性能波动最小，SEA-Guard-12B 的差距低于1分，4B 和 8B 变体也有类似的小差距，展现了强大的跨语言鲁棒性。

我们进一步观察到，在翻译数据集上训练的模型（如 PolyGuard）或缺乏东南亚特定语言和文化设计的模型（如 LionGuard）在文化基准上表现不佳。这些结果强调了文化基础和广泛多语言支持对于安全防护泛化到东南亚场景的重要性。

表1：SEA-SafeguardBench 上的安全防护性能（AUPRC）：In-the-Wild (ITW) 和 Content Generation (CG) 子集

模型	ITW Cultural (英语)	ITW Cultural (东南亚)	CG Cultural (英语)	CG Cultural (东南亚)	提示分类均值	CG Cultural 回复 (英语)	CG Cultural 回复 (东南亚)	回复分类均值
Google Model Armor	86.6	75.6	40.1	33.8	59.0	69.4	59.1	64.2

模型	ITW Cultural (英语)	ITW Cultural (东南亚)	CG Cultural (英语)	CG Cultural (东南亚)	提示分类均值	CG Cultural 回复 (英语)	CG Cultural 回复 (东南亚)	回复分类均值
Azure AI Content Safety	88.5	83.1	37.6	30.2	59.8	-	-	-
OpenAI Moderation	95.3	86.4	45.5	40.3	66.9	-	-	-
LakeraGuard	88.9	76.6	30.0	37.8	58.3	-	-	-
ShieldGemma 2B	95.8	90.6	53.2	51.8	72.8	51.5	47.3	49.4
ShieldGemma 9B	97.2	95.3	52.2	55.7	75.1	56.5	54.0	55.2
ShieldGemma 27B	98.0	96.0	58.7	59.4	78.0	62.8	58.2	60.5
LlamaGuard-3 1B	91.8	86.4	45.7	33.9	64.4	58.6	48.6	53.6
LlamaGuard-3 8B	97.4	95.6	55.4	44.1	73.1	68.0	65.2	66.6
LlamaGuard-4 12B	94.6	84.7	46.0	32.4	64.4	60.9	53.6	57.2
PolyGuard-Qwen 0.5B	97.5	82.6	40.8	32.4	63.3	53.9	43.7	48.8
PolyGuard-Qwen 8B	98.6	94.9	53.8	41.0	72.1	67.9	61.4	64.7
PolyGuard-Ministral 8B	98.9	95.5	49.9	41.1	71.4	64.4	56.2	60.3
Qwen3Guard-Gen 4B	98.4	97.3	56.8	49.0	75.4	72.5	67.7	70.1

模型	ITW Cultural (英语)	ITW Cultural (东南亚)	CG Cultural (英语)	CG Cultural (东南亚)	提示分类均值	CG Cultural 回复 (英语)	CG Cultural 回复 (东南亚)	回复分类均值
Qwen3Guard-Gen 8B	98.7	98.0	54.2	47.6	74.6	74.4	71.1	72.8
LionGuard-2	95.8	78.5	46.7	41.9	65.7	47.8	40.3	44.0
X-Guard	97.0	86.1	42.5	35.1	65.2	-	-	-
SEA-Guard-4B	**99.3**	**98.8**	**58.3**	**61.2**	**79.4**	**73.7**	**69.4**	**71.6**
SEA-Guard-8B	**99.2**	**98.6**	**61.2**	**59.0**	**79.5**	**74.4**	**71.3**	**72.9**
SEA-Guard-12B	**99.5**	**99.0**	**59.7**	**61.7**	**80.0**	**75.4**	**73.2**	**74.3**

4.2 通用安全结果

我们还在英语和东南亚语言的通用安全基准上评估了 SEA-Guard 的性能。与利用通用安全数据集的先前模型（如 PolyGuard）不同，我们的模型未在任何通用数据集上训练；因此，此实验在域外（out-of-domain）设置下解决 RQ1 和 RQ3。

如表2所示，尽管未在通用安全数据上训练，SEA-Guard 泛化良好。SEA-Guard-12B 在提示分类上优于 Qwen3Guard-Gen 8B，在回复分类上仅有0.6分的差距。在所有东南亚语言中，SEA-Guard-12B 在提示分类上始终优于 Qwen3Guard-Gen 8B。

虽然加入通用安全数据集可以提高在通用基准上的性能，但我们的初步实验揭示了一个权衡：添加此类数据会使训练分布偏向通用安全话题，并降低在文化安全内容上的性能——而后者是 SEA-Guard 的主要目标。

> **表2**：通用安全内容上的安全防护性能（AUPRC）。

4.3 零样本视觉-文本安全结果

为解决 RQ3，我们在视觉-文本安全基准上将 SEA-Guard 与视觉-语言模型进行评估。所有模型均在零样本条件下评估，未在视觉安全数据上训练。由于表1中的模型是纯文本的，我们将 SEA-Guard 与支持视觉输入的 LLM 进行比较。

如表3所示，SEA-Guard 取得了一致的改进，在七个设置中的六个上优于竞争模型，仅在 VLGuard 的回复分类上例外。SEA-Guard-4B 和 -8B 在 MSSBench-Embodied 上表现特别好，其家庭任务指令和安全/不安全视觉上下文与我们训练数据中以规范和生活方式为重点的设计高度一致。

相比之下，SEA-Guard-12B 的表现相对较弱，主要是由于其较弱的基础模型（Gemma3-12B-IT）限制了相比 Qwen 和 Qwen-SEA-LION 的提升。尽管如此，SEA-Guard-12B 在所有基准上始终超过 Gemma3-12B 和 Qwen-SEA-LION-v4-8B-VL。

总体而言，这些结果表明，纯文本监督可以诱导出涌现的零样本视觉-文本安全能力，即使 SEA-Guard 主要作为文本安全防护进行优化，也能实现可靠的性能。

> **表3**：视觉-文本安全基准（AUPRC）。p/r 分别代表提示/回复性能。

5 分析

本节通过三个方面研究 SEA-Guard 的有效性：（i）人类对齐分数，（ii）对抗攻击，（iii）数据去重。

5.1 人类对齐

我们评估了模型预测的有害性分数（有害类别的概率）与 SEA-SafeguardBench CG Cultural 子集中人类软标签标注之间的对齐程度。每个样本包含硬标签（安全、敏感、有害）和连续范围[0, 1]内的软标签，该范围被分为三个等间隔，与硬标签类别对齐。

理想情况下，安全防护应跟踪人类判断的严重性，捕捉正确的排序和概率对齐；偏差可能导致系统性的过度防御或防御不足。对齐程度使用 Spearman 和 Pearson 相关系数量化。

如图3所示，SEA-Guard 模型在各严重性级别上实现了更高的 Spearman 和 Pearson 分数以及更清晰的分离，而 Qwen3Guard、LlamaGuard 和 ShieldGemma 则表现出大量重叠。这种在高严重性级别上的防御不足行为构成部署风险，因为有害内容可能绕过安全防护。

处理中间严重性区间对所有模型仍然具有挑战性；它对应于既非明显安全也非明显有害的敏感案例，其处理取决于用户定义的阈值。虽然 SEA-Guard 改善了该区域的分离度，但与相邻区间的区分度不足仍限制了可靠的校准。

> **图3**：模型预测的有害性分数与人类判断的严重性级别之间的对齐。

5.2 对抗攻击鲁棒性

图4显示了安全防护在对抗攻击下对 SEA-SafeguardBench 的鲁棒性，这些攻击保持有害意图但试图逃避检测。我们使用与语言无关的空白字符插入攻击，因为大多数方法（Hughes et al., 2024; Chao et al., 2024b; Jiang et al., 2024）依赖于英语特定的改写或词汇替换，可能无法在非拉丁文字中保持有害意图。

空白字符扰动降低了各模型的预测有害性，表明最小的表面层级变化也能影响安全防护行为。Qwen3Guard-Gen 8B 随扰动强度增加而单调退化，而 LlamaGuard-3 8B 表现出非单调响应，在 K=16 时部分恢复，可能是由于分词器效应。

相比之下，SEA-Guard 模型保持了更强的鲁棒性，在扰动下维持了较高的有害性分数，较大的变体表现出最稳定的分布。

> **图4**：对抗攻击鲁棒性。

5.3 数据集大小和去重研究

图5考察了每种东南亚语言训练数据规模对安全防护性能的影响。从20万到60万样本，性能并非单调增长，表明在中间规模上存在收益递减和潜在的噪声累积。在100万样本时出现了显著提升，表明需要足够大且多样化的数据才能实现规模效益。

值得注意的是，去重后的数据集尽管样本更少，但与完整的100万设置相比实现了相当的性能。虽然20万设置产生了有竞争力的平均 AUPRC，但较小的数据集在稀有、文化特定和对抗性案例上

覆盖不足。因此，我们采用更大规模和去重的数据集，以优先考虑鲁棒性和覆盖率，而非在较小规模上优化平均性能。

> **图5**：数据集大小和去重对模型性能的影响。

6 相关工作

6.1 安全防护模型

先前工作通过使用合成安全数据集调整现有 LLM 来构建多语言安全防护，这些数据集通过多语言提示（Yang et al., 2024; Deng et al., 2025; Joshi et al., 2025）、推理（Liu et al., 2025; Yang et al., 2025）或英语翻译（Upadhayay and Behzadan, 2025; Kumar et al., 2025; Verma et al., 2025）生成。然而，这些方法对东南亚语言的探索仍然很少——这些语言资源匮乏且许多 LLM 对其支持不足。

最近面向东南亚的努力通常依赖翻译或弱监督数据：SEALGuard（Shan et al., 2025）使用 Google 翻译数据，而 LionGuard-2（Tan et al., 2025）在人类聊天数据集上训练轻量级检测器。这些策略——用文化关键词提示或翻译英语数据——缺乏文化基础和质量控制，导致在东南亚文化基准上表现不佳（Tasawong et al., 2025b）。

6.2 文化模型和数据集

先前工作提出了文化主题的数据生成和聚合框架（Li et al., 2024; Thakur et al., 2024; Zhang et al., 2025; Yue et al., 2025; Nyandwi et al., 2025; Feng et al., 2025），但这些工作主要关注使用 GPT-4 等 LLM 的高资源语言，东南亚语言在很大程度上未被探索。

最近面向东南亚的数据集——包括人工标注和合成数据——已开始填补这一空白（Lovenia et al., 2024; Cahyawijaya et al., 2025; Nguyen et al., 2024; Ng et al., 2025），提高了在东南亚基准上的鲁棒性和文化理解（Susanto et al., 2025）。这些研究强调了由于东南亚语言在 LLM 中代表不足，精心设计合成数据的必要性。

7 结论

本文提出了 SEA-Guard——一个支持8种语言、提供三种大小（4B、8B 和 12B）的东南亚区域安全防护模型。该模型在专为东南亚场景设计的新颖数据合成框架上训练，确保数据质量和正确性，以在东南亚语言和文化基准上实现泛化结果。

结果表明，SEA-Guard 在文化安全基准上达到了 SOTA 水平，同时在零样本设置下的视觉-文本基准上优于其他模型。此外，我们的分析还验证了模型在人类对齐、对抗攻击和数据去重方面的鲁棒性。

局限性

虽然我们的模型支持8种东南亚语言（英语也是东南亚的官方语言之一），但仍有一些语言未被覆盖（即高棉语、老挝语、泰卢固语以及超过700种东南亚方言和语言）。这是因为这些语言缺乏可用的基准测试。当新的基准测试可用并支持这些语言时，我们可以轻松扩展模型以出于安全原因支持它们。我们希望向社区强调这一问题——需要安全评估基准，东南亚需要更多关注和努力。

此外，我们承认未在最小的 0.5B 模型上进行实验。我们注意到 0.5B 的性能不可靠，不应用于安全目的，因为模型容易防御不足（即不将任何样本分类为有害），如表1所示 Qwen 0.5B 表现最差。安全至关重要，需要谨慎考虑，因此我们未在 0.5B（Qwen2.5）或 0.6B（Qwen3）等泛化能力不足的模型上进行实验。

伦理声明

关于标注者详情：我们雇用了32名以东南亚语言为母语的标注者（研究生）。我们有4名缅甸语、2名菲律宾语、10名印尼语、4名马来语、6名泰米尔语、2名泰语和4名越南语标注者，每人需审查100个样本/语言。

我们首先进行标注实验，仅选择通过标注测试（即英语测试和安全文本理解测试）的标注者，以测试标注者是否理解并能以高质量方式完成工作。此外，每位标注者的薪酬为18美元/小时，高于平均薪酬水平。我们也请标注者在标注前考虑数据的敏感性，因为数据集中的某些样本可能对他们来说过于敏感。标注者可以自由选择退出，如果他们对此过程感到不适。

关于我们工作中的潜在风险：我们承认生成的数据集包含不安全样本的有害内容。然而，我们数据集和模型的目的和用途是对输入的安全性进行分类，而非训练任何 LLM 生成有害内容。我们鼓励所有未来使用我们工作的研究人员和个人不要使用我们的数据集来生成更多有害内容。

参考文献

- Azure (2025). Azure ai content safety documentation.
- Cahyawijaya et al. (2025). Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In ACL 2025.
- Chao et al. (2024a). Jailbreakbench: An open robustness benchmark for jailbreaking large language models. arXiv:2404.01318.
- Chao et al. (2024b). Jailbreaking black box large language models in twenty queries. arXiv:2310.08419.
- Deng et al. (2025). Duoguard: A two-player rl-driven framework for multilingual llm guardrails. arXiv:2502.05163.
- Evert (2004). The statistics of word cooccurrences: Word pairs and collocations.
- Feng et al. (2025). CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In ACL 2025.
- Geng et al. (2025). VSCBench: Bridging the gap in vision-language model safety calibration. In Findings of ACL 2025.
- Google Cloud (2025). Model armor overview.
- Han et al. (2024). Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. arXiv:2406.18495.
- Hughes et al. (2024). Best-of-n jailbreaking. arXiv:2412.03556.
- Inan et al. (2023). Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv:2312.06674.
- Ji et al. (2025). Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. arXiv:2406.15513.
- Jiang et al. (2024). Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. arXiv:2406.18510.

- Joshi et al. (2025). Cultureguard: Towards culturally-aware dataset and guard model for multilingual safety applications. arXiv:2508.01710.
- Kumar et al. (2025). Polyguard: A multilingual safety moderation tool for 17 languages. In Second Conference on Language Modeling.
- LakeraAI (2025). Lakeraguard.
- Li et al. (2024). Culturellm: Incorporating cultural differences into large language models. In NeurIPS 2024.
- Liu et al. (2025). Guardreasoner: Towards reasoning-based LLM safeguards. In ICLR 2025 Workshop.
- Lovenia et al. (2024). SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In EMNLP 2024.
- Ng et al. (2025). Sea-lion: Southeast asian languages in one network. arXiv:2504.05747.
- Nguyen et al. (2024). SeaLLMs - large language models for Southeast Asia. In ACL 2024 System Demonstrations.
- Nyandwi et al. (2025). Grounding multilingual multimodal LLMs with cultural knowledge. In EMNLP 2025.
- OpenAI (2024). Upgrading the moderation api with our new multi-modal moderation model.
- Ren and Xiong (2023). HuaSLIM: Human attention motivated shortcut learning identification and mitigation for large language models. In Findings of ACL 2023.
- Shan et al. (2025). Sealguard: Safeguarding the multilingual conversations in southeast asian languages for llm software systems. arXiv:2507.08898.
- Susanto et al. (2025). SEA-HELM: Southeast Asian holistic evaluation of language models. In Findings of ACL 2025.
- Tan et al. (2025). LionGuard 2: Building lightweight, data-efficient & localised multilingual content moderators. In EMNLP 2025 System Demonstrations.
- Tasawong et al. (2025a). Shortcut learning in safety: The impact of keyword bias in safeguards. In LLM Security Workshop (LLMSEC).
- Tasawong et al. (2025b). Sea-safeguardbench: Evaluating ai safety in sea languages and cultures. arXiv:2512.05501.

- Thakur et al. (2024). Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In NAACL 2024.
- Upadhayay and Behzadan (2025). X-guard: Multilingual guard agent for content moderation. In LLM Security Workshop (LLMSEC).
- Verma et al. (2025). MULTIGUARD: An efficient approach for AI safety moderation across languages and modalities. In EMNLP 2025.
- Wang et al. (2022). Identifying and mitigating spurious correlations for improving robustness in NLP models. In Findings of NAACL 2022.
- Wei et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS 2022.
- Yang et al. (2025). MrGuard: A multilingual reasoning guardrail for universal LLM safety. In EMNLP 2025.
- Yang et al. (2024). Benchmarking llm guardrails in handling multilingual toxicity. arXiv:2410.22153.
- Ye et al. (2025). The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. arXiv:2402.12715.
- Yue et al. (2025). Pangea: A fully open multilingual multimodal LLM for 39 languages. In ICLR 2025.
- Zeng et al. (2024). Shieldgemma: Generative ai content moderation based on gemma. arXiv:2407.21772.
- Zhang et al. (2025). CultureSynth: A hierarchical taxonomy-guided and retrieval-augmented framework for cultural question-answer synthesis. In Findings of EMNLP 2025.
- Zhao et al. (2025). Qwen3Guard. (referenced in paper)
- Zhou et al. (2025). MSSBench. (referenced in paper)
- Zong et al. (2024). VLGuard. (referenced in paper)