统一游戏监管：软提示和大语言模型辅助标签转移以实现资源高效的毒性检测

Zachary Yang
zachary.yang@mail.mcgill.ca
Ubisoft La Forge | McGill University | Mila
Montreal, Canada

Domenico Tullo
Ubisoft La Forge
Montreal, Canada

Reihaneh Rabbany
McGill University | Mila | CIFAR AI Chair
Montreal, Canada

Abstract

在游戏社区中进行有害行为检测面临显著的扩展挑战，尤其是在涉及多个游戏和语言的实时环境中，这使得计算效率变得至关重要。我们提出了两个关键发现来应对这些挑战，并基于我们先前关于 ToxBuster 的工作进行改进，这是一种基于 BERT 的实时有害行为检测系统。首先，我们介绍了一种软提示方法，通过引入游戏上下文标记，使单个模型能够有效处理多种游戏，其性能与课程学习等更复杂的方法相匹配，并提供卓越的可扩展性。其次，我们开发了一种基于 GPT-4o-mini 的 LLM 辅助标签转移框架，以扩展对另外七种语言的支持。在实际游戏聊天数据的评估中，法语、德语、葡萄牙语和俄语的宏 F1 评分从 32.96 % 到 58.88 % 不等，其中德语表现尤为突出，超过了英语基准的 45.39 %。在实际应用中，与维护每个游戏和语言组合的独立模型相比，这种统一方法显著减少了计算资源和维护开销。在育碧，该模型成功识别出每个游戏每天平均 50 名玩家从事可处罚行为。

CCS Concepts

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; *Information extraction*; • **Applied computing** → **Sociology**.

Keywords

Toxicity, Chat Moderation, Scaling, Soft-Prompting, LLM-Assisted Label Transfer

ACM Reference Format:

Zachary Yang, Domenico Tullo, and Reihaneh Rabbany. 2025. 统一游戏监管：软提示和大语言模型辅助标签转移以实现资源高效的毒性检测. In *Proceedings of Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3711896.3737271

1 介绍

在线毒性已成为数字平台上的一个普遍挑战，游戏社区尤其受到严重影响。反诽谤联盟（ADL）报告称，截至 2023 年，76 % 的成年人和 75 % 的青少年和儿童在游戏空间中遭受骚扰。除了记录在案的心理伤害和可能引发现实世界暴力的潜力 [**adl_2021**] 之外，毒性行为直接影响着游戏公司的收入来源，有 20 % 的玩家因骚扰经历减少了支出。这个挑战涉及多个平台，从社交媒体（Facebook [**hate_speech_on_facebook**] 、Reddit [**impact_of_toxic_language_on_reddit**] 、YouTube [**hate_speech_in_youtub**）到各种游戏环境 [**playing_against_hate_speech**, **yang-etal-2023-towards-**。为了应对这个挑战，我们之前的工作推出了 ToxBuster [**yang-etal-2023-towa**，这是一个为生产部署而设计的实时聊天毒性检测模型。然而，将这一解决方案扩展到育碧多样化的游戏组合中揭示了两个关键挑战：需要将特定游戏的模型统一为一个可部署的解决方案，并需要支持多种语言。这些挑战反映了更广泛的行业对可扩展的多语言内容审核系统的需求。第一个挑战源于为每个游戏维护单独模型的操作复杂性。虽然我们最初创建特定游戏模型的方法展示了其效果，但随着游戏组合的扩大，这种方法证明是不可持续的。在探索了几种统一方法，包括混合数据集训练和课程学习后，我们成功地采用了软提示技术，在显著提高可扩展性的同时保持性能。第二个挑战涉及在语言间扩展有害内容检测，同时解决不一致的有害内容定义的基本问题。尽管研究关注不断增加，该领域仍缺乏标准化定义 [**https://doi.org/10.48550/arxiv.1809.07572**] ，这导致任务制定的各种方法 [**Vidgen2020**] 。当组织需要随时间更新有害内容类别时，这一挑战尤其严重，正如我们在育碧的经历。为了解决这一问题，我们开发了一个创新的 LLM 辅助标签转移框架，该框架利用现有数据集，同时适应特定组织的定义。我们将这一框架应用于 15 个开源有害内容数据集，涵盖中文、法语、德语、日语、葡萄牙语和俄语，创建了一个综合的多语言社交网络有害内容（MLSNT）数据集，与我们当前的有害内容定义一致。

总结起来，我们的贡献如下：

(1) 一种软提示方法，能够在保持性能和提高可扩展性的同时，成功统一针对特定游戏的恶意语言模型
(2) 通过我们的 LLM 辅助标签转移框架创建的新的多语言毒性数据集 MLSNT
(3) 一个统一的、可用于生产的毒性检测模型，能够有效处理多种游戏和语言

可重复性：数据集发布为 复杂数据实验室/机器学习和网络科学小组 。

2 相关工作

过去十年中，在线平台的毒性检测引起了显著关注。早期的研究将毒性检测构建为一个简单的分类任务，采用带有人工特征的传统机器学习方法 [**hate_speech_on_twitter**] 。随后的方法利用深度神经网络来更好地捕捉文本中的上下文和细微差别 [**gamback-sikdar-2017-using**, **content_driven_detection**] ，而最近的研究则采用预训练语言模型来利用其上下文化的表示 [**ALMEREKHI2022100019**, **DBLP:journals/corr/abs-2201-00598**, **perspective_api**] 。在需要快速和上下文感知的响应的在线游

KDD '25, August 3–7, 2025, Toronto, ON, Canada.

ACM ISBN 979-8-4007-1454-2/25/08
https://doi.org/10.1145/3711896.3737271

戏环境中，研究者开发了专门的检测系统 [**10.1145/2998181.2998213**, **yang-etal-2023-towards-detecting**, **10.1145/3675805**] 。在线交流的全球性进一步激发了对多模态和多语言毒性检测的研究，相关工作通过整合跨语言转移、视觉线索和音频信号，旨在稳健地解决多元社区中的毒性问题 [**zampieri-etal-2019-semeval**, **bui2024multi3hatemultimodalmultilingualmulticultural**, **cao-etal-2024-toxicity**] 。

把语言模型（用于毒性检测）扩展到处理大型、异质数据集的规模，需要探索多样的训练范式。虽然早期的方法使用端到端的单任务目标训练，但最近的研究强调了混合训练机制的好处，这种机制结合了有监督的微调与无监督的预训练 [**kaplan2020scalinglawsneurallanguage**, **brown2020languagemodelsfewshotlearners**, **ke2023continualpretraininglanguagemodels**] 。我们的工作结合了基础模型的持续预训练和领域适应 [**farahani2020briefreviewdomainadaptation**] 。课程学习策略，即模型逐步暴露于难度逐渐增加的示例，已被证明对管理类别不平衡和上下文变化有效 [**10.1145/1553374.1553380**, **soviany2022curriculumlearningsurvey**] 。我们实现了这种方法的一种简单版本。软提示和元学习的相关进步使得大规模预训练模型的微调更加高效，在降低计算开销的同时，在毒性检测任务上保持了强大的性能 [**huang-etal-2023-learning-better**, **li2021prefixtuningoptimizingcontinuousprompts**] 。

大型语言模型的出现已经将毒性检测扩展到了零样本和少样本学习领域。开创性的研究表明，像 GPT-3 [**brown2020languagemodelsfewshotlearners**] 这样的模型能够在极少任务特定监督的情况下，推广检测有害内容。这为基于提示的方法铺平了道路，通过引发结构化推理来改进有毒内容的识别 [**koh2024llmsrecognizetoxicitystructured**, **shaikh-etal-2023-second**, **PAN20242849**, **plaza-del-arco-etal-2023-respectful**] 。尽管这些研究强调了大型语言模型在处理毒性检测方面的多功能性，但偏见、与经过微调的小型语言模型在基准测试中相比表现不佳以及可解释性等挑战仍然不断激发着进一步的研究，正如最近的调查所指出的那样 [**bommasani2022opportunitiesrisksfoundationmodels**, **cao-etal-2024-toxicity**] 。

实时毒性检测对于确保安全的在线互动仍然至关重要，尤其是在社交媒体和游戏等动态环境中。小型语言模型（SLMs）依然是最实用的解决方案，在数据集创建的前期成本与检测性能、推理速度和成本效率 [**yu2023openclosedsmalllanguage**] 之间取得了有吸引力的平衡，相较于大型语言模型的持续成本。这些模型依赖于精简的架构和高效的推理技术，必不可少于资源受限环境的部署。最近的实证研究证明，SLMs 能够实时调节内容而不牺牲准确性，从而在实用性上优于更重、更耗计算资源的替代品 [**yu2023openclosedsmalllanguage**, **zhang2023efficienttoxiccontentdetection**] 。

3 方法论

我们的方法解决了两个关键的生产挑战：将游戏特定的模型统一为一个可部署的模型，以及将毒性检测能力扩展到多种语言。本节详细介绍我们对这两个挑战的方法。

3.1 统一特定游戏的模型

为了满足生产部署的需求，需要将我们针对游戏的毒性检测模型整合为一个能够处理多个游戏的统一模型。之前，ToxBuster 对每个游戏（两个流行的多人游戏）使用单独的模型，这增加了操作的复杂性和资源的使用。我们研究了在简化部署架构的同时保持模型有效性的方法，评估了四种复杂性递增的训练方法：

(1) 单独游戏（基线）：在特定游戏数据集上微调的单个模型，代表我们之前的工作。在生产中，我们会选择表现最好的模型。
(2) 混合数据集：一种统一的方法，将所有游戏的数据组合成一个数据集，以微调一个模型，从而显著简化部署。
(3) 课程学习：一种混合数据集方法的扩展，在这种方法中，我们依次在单个游戏数据集上继续进行微调。虽然这会产生多个中间模型，但我们选择在所有游戏中表现最好的模型进行部署。
(4) 软提示：一种维持单一模型的方法，同时通过专用输入标记保留特定游戏知识。我们为每个输入序列加上一个 GAME_TYPE_TOKEN，具有三个可能的值：GAME_1、GAME_2 或 GAME_UNKNOWN。这使得模型能够根据游戏上下文调整其行为。尽管 GAME_UNKNOWN 标记没有经过训练，但可以在推理时使用，例如当游戏未知（由于 API/生产限制）或测试新游戏时。

所有实验都对 ToxBuster 进行了微调——使用 bert-base-uncased 模型，最大长度为 512 个标记，结合了聊天上下文，直至达到此限制。为了确保评估的稳健性，我们进行了五次随机种子不同的实验。对于软提示实验，我们评估了游戏感知（GAME_x 标记）和与游戏无关（GAME_UNKNOWN 标记）的场景，以模拟真实世界中的部署条件。性能使用每个游戏数据集的宏 F1-score 以及各个游戏的总体平均分进行测量。

3.2 多语言扩展

为了扩展 ToxBuster 在英语之外的能力，我们评估了几种预训练的多语言 BERT 变体，以替代我们现有的基于 BERT 的模型。我们使用 Game_1 数据作为基准，对比了四种架构及其多语言对应版本（BERT，DistilBERT，DeBERTa 和 RoBERTa）。基于三个标准，XLM-RoBERTa-base 成为我们最佳选择：卓越的宏 F1 分数表现，简化的连续预训练过程（只需掩码语言建模而不是同时进行 MLM 和下一句预测），以及其使用全面的 CC-100 数据集进行预训练。

3.2.1 **语言识别**. 我们使用 Lingua-py [1] 分析了游戏聊天数据中的语言分布，以优先考虑我们的多语言扩展。虽然这个工具在文本长度较短时的准确性降低——这是与通常简短的游戏聊天消息相关的一个限制——但它为我们的初步语言评估提供了足够的见解。我们分析了两款游戏一周的聊天数据，着重关注于与 CC-100 数据集覆盖、Lingua 的能力以及我们全球玩家基础相交的 12 种语言：英语、西班牙语、法语、德语、俄语、葡萄牙语、日语、韩语、中文、阿拉伯语、印地语和泰语。我们的分析表明，英语在聊天数据中占据主导地位，占比为 80 %，紧随其后的是法语、德语、日语、葡萄牙语和俄语，这些是接下来最常见的语言。

3.2.2 *LLM* **辅助标签转换**. 毒性检测的一个基本挑战在于数据集之间定义的不一致性。即使在单个组织内部，随着业务规则和法律法规的变化，毒性类别也在演变，导致定义随着时间的推移而被更新、拆分或合并。尽管许多现有数据集可能有潜在的利用价值，但获取专家审查以实现标签标准化既具有挑战性又昂贵。为此，我们提出了一种基于大型语言模型（LLM）协助的标签转移方法。我们的方法基于三个关键假设：

(1) 现有数据集（无论是专有的还是开源的）中的原始标签被视为真实值，因为它们是由人工标注的。

[1] https://github.com/pemistahl/lingua-py

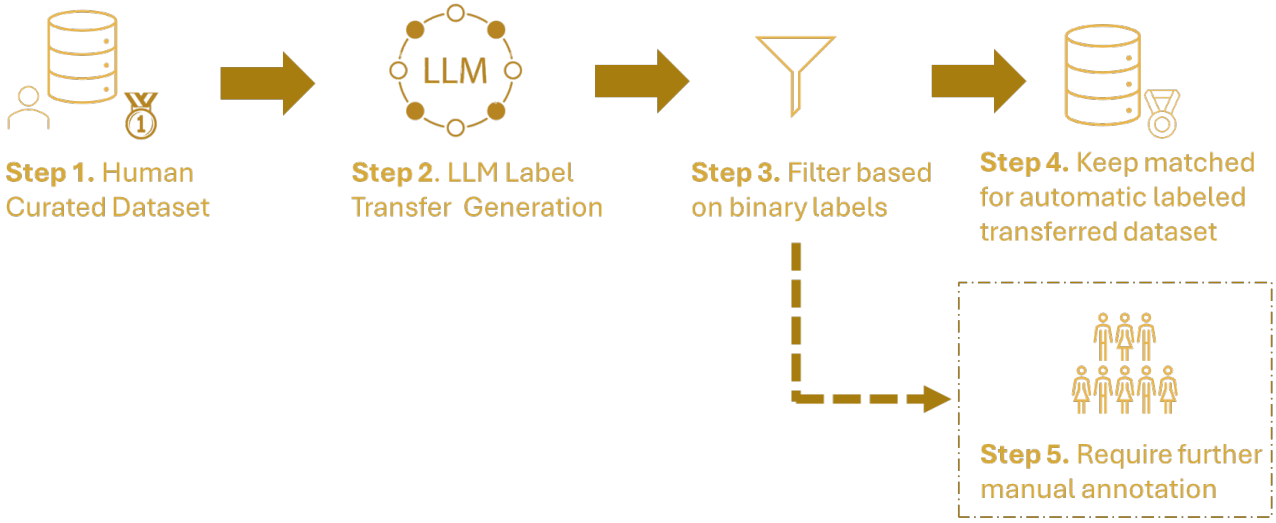

Figure 1: 用于跨数据集有害内容分类的 LLM 辅助标签转移框架。该框架通过基于 LLM 的标签生成利用现有的人类整理的数据集，然后通过验证过滤器确保人类和 LLM 注释之间的一致性。

(2) 在数据集之间的二元分类中，有足够的一致性——源数据集中标记为有害或无害的内容与我们的用例定义一致。
(3) 对于细粒度的有害类别，我们可以接受较低的性能（例如，准确率，F1-分数）。

我们通过如下工作流程来实现标签转移，如图 1 所示：

(1) 获取一个与毒性检测相关的人类标注数据集
(2) 将原始的人类标注转换为二元分类
(3) 使用一个大型语言模型（例如，GPT-4o-mini）来生成二元（有毒/无毒）和具体类别标签
(4) 仅保留人工和 LLM 二分类一致的条目

我们评估了 GPT-4o-mini 对齐我们毒性分类框架的能力，使用的提示结合了元提示 [**zhang2024metapromptingaisystems**]、思维链 [**wei2023chainofthoughtpromptingelicitsreasoning**] 和自我一致性 [**wang2023selfconsistencyimproveschainthought**]。模型的任务是提供二元分类（有毒/无毒），识别有毒跨度，并分配特定的毒性类别。我们使用我们的黄金数据集验证了这种方法，该数据集由三个注释者一致同意的条目组成。由于某些限制，我们当前的实现到此为止。我们计划的下一步是让审核员查看之前人工注释者和大型语言模型标签不一致的被丢弃案例。这种方法的关键优势在于能够有效收集非英语语言中明确的有毒和无毒案例，保留宝贵的人类注释资源，用于需要细致判断的模糊案例。

3.2.3 **人工注释的数据集**. 我们收集了来自 hatespeechdata.com 和其他同行评审出版物的人为标注仇恨言论数据集，涵盖七种语言：法语（1 个数据集 [**ousidhoum-etal-2019-multilingual**]）、德语（3 个数据集 [**goldzycher-etal-2024-improving**, **germ_eval**, **10.1145/3368567.3368584**]）、葡萄牙语（3 个数据集 [**brasnam**, **10.1007/s10579-023-09657-0**, **leite-etal-2020-toxic**]）、俄语（2 个数据集 [**DBLP:conf/raslan/AndrusyakRK18**, **saitov-derczynski-2021-abusive**]）、简体中文（3 个数据集 [**deng-etal-2022-cold**, **JIANG2022100182**, **lu-etal-2023-facilitating**]）、繁体中文（1 个数据集 [**10.1109/IRI51335.2021.00069**]）和日语（2 个数据集 [**DBLP:journals/corr/abs-2407-03963**]）。表 9 详细说明了每个数据集的原始分类任务、来源平台、数据量以及用于 GPT-4o-mini 辅助标签转移的估计计算成本。我们应用了我们的 LLM 辅助标签转移框架，将这 15 个数据集转换为我们的毒性类别，形成了社交网络系统 (SNS) 数据集。

3.2.4 *ToxBuster*. 我们通过评估三个基础模型来改造 ToxBuster，使其具备多语言能力：bert-base（基线）、xlm-roberta-base 和 xlm-roberta-base-adapted。xlm-roberta-base-adapted 模型在一个星期的来自两个游戏的聊天数据上进行了 74 个 epoch 的预训练，当遮蔽语言模型损失趋于稳定时提前停止（最多 100 个 epoch）。我们应用了 GAME_TYPE_TOKEN 值的软提示：GAME_1、GAME_2、MLSNT 和 GAME_UNKNOWN。性能通过加权和宏观 F1-分数进行测量。

3.2.5 **人类评估**. 为了评估 MLSNT 在这两个游戏上的迁移能力，我们为每个游戏和语言开发了一个系统的采样策略。我们针对每个类别选取 50 个样本，并设计了一个 20 % 的溢出机制用于转入下一个有害类别以及 80 % 的机制用于转入非有害类别，保持类似于英语的有害失衡。通过动态计算可用样本，并根据不足进行目标调整，该方法确保有害和非有害内容更加平衡地呈现。一名在每种语言上都很精通并熟悉游戏的人工标注者验证了他们各自的 450 行评估集。

4 结果

我们将实验结果分为两个部分：首先，我们评估了我们的方法如何在多个游戏中扩展单个模型，接着介绍我们将模型扩展以支持七种语言的结果。

4.1 统一游戏特定模型

在表 1 中，我们使用宏观标记级别的 F1 分数比较了在 GAME_1 和 GAME_2 数据集上的四种不同方法。整体得分表示在两场游戏上平衡的 F1 分数，是五次运行的平均值。

单一游戏模型，作为基线，在 GAME_1 上实现了 42.84 % 的宏观 F1 得分，在 GAME_2 上实现了 38.80 % 的宏观 F1 得分。通过混合训练结合数据集将整体性能提高到 42.48 %，表明混合

Table 1: 单项、混合、课程学习以及软提示的宏观 F1 分数。对于单项游戏，我们在与测试集相同的数据集上进行训练。对于其余的，我们在括号中显示训练数据集，其中 "&" 表示数据集的混合，"|" 分隔训练步骤。 在这里，我们评估软提示，在推理时不提供游戏。即使在推理期间没有明确的游戏信息，软提示的表现也与课程学习相当，表明其在跨游戏泛化方面的有效性。**

Method		G_1	G_2	Overall
Single	(Game_1) or (Game_2)	42.84 ± 2.56	38.80 ± 0.53	40.82 ± 0.87
Mixed	(G_1 & G_2)	45.10 ± 0.73	39.87 ± 2.09	42.48 ± 0.74
Curriculum	(G_1 & G_2 | G_1)	46.32 ± 0.40	40.37 ± 2.12	43.35 ± 1.16
	(G_1 & G_2 | G_2)	44.19 ± 1.82	40.23 ± 2.18	42.21 ± 1.52
Soft Prompting	(G_1 & G_2)	45.39 ± 1.01	40.94 ± 1.38	43.16 ± 1.06
	(G_1 & G_2)**	45.24 ± 1.13	40.82 ± 1.46	43.03 ± 1.19

数据集实际上对两个游戏都有益（提高了 1.7 %）。课程学习在（GAME_1 & GAME_2 | GAME_1）序列中实现了最高的 43.35 % 的宏观 F1 得分，但这一方法存在显著的可扩展性限制——每增加一个新游戏都需要在所有游戏的组合数据集上进行训练，然后在特定游戏的数据集上进行训练，并进行广泛的模型选择以优化整体性能。

软提示成为一种更实用的解决方案，当推理中提供游戏类型时，它实现了第二高的宏观 F1 分数，而当未提供时，它实现了第三高的分数。这种方法提供了优越的可扩展性和效率优势：

(1) 只需要在组合的多游戏数据集上进行一次训练，与课程学习相比，可以大大减少训练时间
(2) 支持新游戏只需引入一个新的 GAME_TYPE_TOKEN 并重新训练。

Table 2: GAME_TYPE_TOKEN 位置对软提示的影响。GAME_TYPE_TOKEN 需要位于 CONTEXT 部分之前。

Method		Overall
Mixed		42.48 ± 0.74
Soft Prompting (Before Context)	(GAME_1 & GAME_2)	43.16 ± 1.06
	(GAME_1 & GAME_2) **	43.03 ± 1.19
Soft Prompting (Before Current Line)	(GAME_1 & GAME_2)	41.84 ± 1.32
	(GAME_1 & GAME_2) **	41.57 ± 1.37

4.1.1 **消融**。鉴于我们模型的独特架构，它包含了针对上下文和当前行的独立模块，我们进行了一个消融研究，以确定 GAME_TYPE_TOKEN 的最佳放置位置。我们比较了它放置在上下文之前与直接置于当前行之前的效果。正如表 2 所示，GAME_TYPE_TOKEN 在上下文之前放置时效果最佳。如果放在当前行之前，性能实际上会下降。

4.2 多语言扩展

4.2.1 ***LLM* 辅助标签转移的可行性**。在表 3 中，我们评估了在一个所有三个人类标注者完全达成一致的数据集上使用 GPT-4o-mini 进行标签转移的可行性。对于有害类别进行了两项关键修改：（1）将威胁分为威胁生命和不威胁生命的类别，（2）为性内容/骚扰提取了一个特定类别，同时将其他类别合并到潜在有害内。

Table 3: GPT-4o-mini 标签转移。* 代表使用更新的有害类别集的数据集。大型语言模型可以用于自动化不同注释方案之间标签的复杂转移任务。

	Prompt	Temp	Weighted F1-Score Binary	Weighted F1-Score Class-wise	True Labels
Full	v1	1.0	86.66 %	84.48 %	136,605
Full*	v1	0.7	85.44 %	82.23 %	133,951
	v1	1.0	85.35 %	81.89 %	132,009
	v2	0.7	82.51 %	76.85 %	127,392
	v2	1.0	82.47 %	76.79 %	127,305

实验结果表明，GPT-4o-mini 的标签转移功能具有潜力。在原始完整数据集上，该模型在二元有毒/无毒分类任务中达到了 86.66 % 的加权 F1 分数，而加权类别 F1 分数稍低，为 84.48 %。当在更新了有毒类别的数据集上进行测试时，二元加权 F1 分数保持一致，但类别性能略有下降。

为了优化标签转移，我们研究了两种提示词变体（v1 和 v2）以及温度设置（0.7 和 1.0）。鉴于我们的框架侧重于仅保留 "真实" 标签——即人类标注者和大型语言模型一致同意的标签——我们确定提示词版本 1 和温度为 0.7 是在最大化真实标签数量和加权 F1 分数方面的最佳配置。

Table 4: GPT-4o-mini 标签传输过滤。通过限制仅使用人类和 LLM 一致的二元标签，我们在有害类别的性能上提升了大约 40 %。

	Prompt	Temp	No Filter	LLM "Toxic"	Agreed Toxic	Agreed Labels
Full	v1	1.0	84.48 %	38.36 %	79.12 %	98.10 %
Full*	v1	0.7	82.23 %	29.30 %	67.90 %	97.20 %
	v1	1.0	81.89 %	29.55 %	67.29 %	96.97 %
	v2	0.7	76.85 %	22.74 %	61.24 %	95.66 %
	v2	1.0	76.79 %	22.73 %	61.41 %	95.66 %

我们通过研究过滤方法更深入地探讨加权类别 F1 得分。表格 4 展示了我们提出的过滤方法的关键重要性，该方法重点关注人类标注者和 GPT-4o-mini 一致的情况。结果揭示了我们提出的方法中关于仅保留人类标注者和大语言模型一致的情况的关键见解。当考虑未过滤的预测时，基准性能为 84.48 %。然而，当我们仅将分析限制在 GPT-4o-mini 的有害预测时，性能显著低，仅达到 38.36 %。这突显了仅依赖于大型语言模型分类的挑战。当我们过滤出人类标注者和 GPT-4o-mini 都将内容分类为有害案例时，最显著的发现出现了：性能显著提高到 79.12 %，表现出 40 % 的显著提升。这一一致模式在不同的提示版本和温度设置中都成立。因此，我们的方法在需要标签转移时，提供了一种减少人工标注成本的解决方案。从直觉上看，可以有效捕捉明确的有害和无害内容案例，当有毒标签需要重新分配时，提供可靠的自动过滤。

4.2.2 **多语言基础模型搜索**。在表 5 中，我们比较了 BERT、DistilBERT、DeBERTa 和 RoBERTa 模型的英文和多语言版本之间的宏观和加权 F1 分数。我们的分析揭示了几个关键模式。虽然所有多语言版本由于非英语语言的标记词汇扩展而显示出参数数量增加，但它们在计算上仍然高效，参数少于 300M，因

Table 5: 多语言基础 BERT 模型搜索。考虑到训练数据集的时效性和模型性能，优先选择 xlm-roberta-base。

Model	Params	F1-Score	
		Macro	Weighted
bert-base-uncased	110M	43.30 ± 1.73	82.02 ± 0.33
...-multilingual-cased	179M	39.31 ± 2.68	81.28 ± 0.49
distilbert-base-uncased	67M	41.07 ± 2.05	81.46 ± 0.66
...–multilingual-cased	135M	37.49 ± 2.08	80.61 ± 0.61
deberta-v3-base	184M	41.97 ± 3.12	82.54 ± 0.50
mdeberta-v3-base	276M	38.18 ± 4.40	81.71 ± 0.64
roberta-base	125M	40.35 ± 2.09	81.81 ± 0.35
xlm-roberta-base	279M	41.80 ± 2.76	81.56 ± 0.59

此推理时间增加可以忽略不计。向多语言版本的过渡通常导致所有模型的加权 F1 分数下降。BERT、DistilBERT 和 DeBERTa 表现出类似的性能下降模式，宏观 F1 分数下降约为 3-4 %。值得注意的是，XLM-RoBERTa 展示了不同寻常的表现，与仅英语版本相比，宏观 F1 分数略有提高。

4.2.3 **多语言数据集** *(MLSNT)*. 使用提示 v1 在温度 0.7 的条件下，我们在所有 15 个数据集上提示 GPT-4o-mini。表 6 显示了处理后的数据集结果，其中我们展示了处理后剩余的行数（人类和 GPT-4o-mini 都同意的部分），被丢弃行的%，原始和处理后数据集的毒性百分比。一个关键观察是，在处理过程中被丢弃行的百分比变化从 10.02 %（LLM_JP）到 69.35 %（HASOC）。被丢弃的行未显示出与毒性百分比变化的明确相关性。值得注意的是，大多数数据集在处理后显示有害行的百分比增加，特别是在德语（GERM_EVAL：增加 19.94 %，HASOC：增加 20.79 %）和俄语（Abusive：增加 21.27 %）等语言中。毒性百分比在不同语言和数据集之间差异显著，一些如俄语 South Park 和日语 LLM_JP 数据集显示变化很小（分别为-0.14 % 和 0.71 %），而其他则表现出显著的转变。这种变化突显了毒性的细微差别以及语言特异性考虑的重要性。

4.2.4 **多语言** *ToxBuster*. 然后，我们使用多语言基础变换器模型（具体而言，xlm-roberta-base 和 xm-roberta-base-adapted）将 ToxBuster 训练成其多语言变体。我们在表 7 中报告了在我们原始游戏数据集（GAME_1 和 GAME_2）上的宏观和加权 F1 分数、游戏数据集上的总体表现，最后是在 MLSNT 上的表现。

与多语言基础模型搜索结果相似的是，bert-base 在宏 F1 值上（43.17 -> 40.71）优于 roberta-base。

通过使用软提示来包含 MLSNT，我们发现从仅英语到多语言不会影响性能。最后，我们看到 xlm-roberta-base 的领域适应版本 xlm-roberta-base-adapted 相比其基础版本得分略有提高。

4.2.5 **多语言游戏聊天的人类评估**. 使用微调过的 xlm-roberta-base-adapted，我们从每种语言和游戏中抽取出推断的行，每种语言和每个游戏总计 450 行。我们在表 **??** 中报告宏观和加权类别的 F1 分数。这种人工验证揭示了多语言毒性检测的挑战（包括从 MLSNT 转移到游戏聊天和与英语相比资源较少的情况）。

我们看到显著的性能差异，宏观 F1 分数从日产的 GAME_1 的关键低点 19.07 % 到德语 GAME_1 的相对稳健的 58.88 % 不等。我们注意到，对于日语，聊天内容主要是敌对的/从另一种语言翻译而来，并不真正包含正常可理解的日语。这些差异突显了跨语言毒性分类中固有的显著语言和语境复杂性。各语言之间的性能不均匀——例如法语的宏观分数为 45.27 %，权重分数为 82.78 % ——阐明了开发通用多语言毒性检测系统的挑战。这种变化很可能源于语言差异、文化交流规范、特定平台的讨论模式，甚至是在不同语言数据集之间可能存在的注释策略差异。

5 结论

在本研究中，我们解决了工业化游戏聊天系统毒性检测模型的两个基本挑战：统一游戏特定模型和扩展多语言能力。我们的软提示法（soft-prompting approach）表明，多个游戏特定模型可以在不牺牲性能的情况下有效统一，实现了 43.16 % 的宏观 F1 分数，同时显著降低了操作复杂性和训练时间。LLM 辅助的标签转移框架不仅实现了高效的跨语言毒性分类，还为在组织需求演变时适应毒性定义提供了可扩展的解决方案。更具体地说，通过整合 XLM-RoBERTa 和新的 MLSNT（多语言社交网络毒性）数据集，该数据集跨越七种语言的 15 个来源，我们开发了一种可投入生产的多语言 ToxBuster，而不会牺牲其在英语数据集上的性能。对真实游戏聊天数据的人类评估揭示了不同语言之间显著的性能差异（宏观 F1 分数从 19.07 % 到 58.88 %），其中一些甚至超过了英文基准 45.39 %。

6 局限性

我们的工作在多游戏、多语言的毒性检测方面取得了显著的进展，同时也指出了未来研究的几个领域。我们的评估聚焦于两款热门的育碧游戏——代表了不同游戏类型和独特的玩家互动——尽管未来的工作可以探索更多的游戏类型。通过 LLM 辅助标签转移方法成功地减少了人工标注的需求，尽管不同语言的标注需求有所不同（10-70 %），这表明还有进一步优化的机会。我们的人工评估显示了不同语言之间的性能差异，尤其是对于那些训练数据较少或书写系统复杂的语言，强调了持续进行数据收集努力的重要性。在线交流的动态性带来了固有的挑战，因为表达模式随着时间推移和文化背景而变异。这些观察结果与内容审核领域的更广泛行业挑战相一致，并为跨文化毒性检测领域的未来研究指明了有前景的方向。

Acknowledgments

We wish to thank Ubisoft La Forge, Ubisoft Montreal User Research Lab and Ubisoft Data Office for providing technical support and insightful comments on this work. We also acknowledge funding in support of this work from Ubisoft, the Canadian Institute for Advanced Research (CIFAR AI Chair Program) and Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship-Doctoral (PGS D) Award.

A 附录

Received 10 February 2025; accepted 22 May 2025

Table 6: MLSNT 数据集概览。被丢弃的行比例范围为 10-70%。在大多数情况下，数据集中的有害行比例增加。

Language	Name	Lines		Toxicity %		
		Processed	% Discarded	Original	Processed	Δ
Chinese (Simplified)	COLD [**deng-etal-2022-cold**]	20,087	46.18 %	48.03 %	60.67 %	12.64 %
Chinese (Simplified)	SWSR [**JIANG2022100182**]	5,708	36.32 %	34.50 %	47.85 %	13.35 %
Chinese (Simplified)	TOXICN [**lu-etal-2023-facilitating**]	8,500	29.23 %	53.79 %	56.51 %	2.71 %
Chinese (Traditional)	TOCAB [**10.1109/IRI51335.2021.00069**]	65,263	37.25 %	14.48 %	8.94 %	-5.54 %
French	MLMA [**ousidhoum-etal-2019-multilingual**]	3,203	20.20 %	79.55 %	93.57 %	14.02 %
German	GAHD [**goldzycher-etal-2024-improving**]	7,886	28.28 %	42.43 %	55.88 %	13.45 %
German	GERM_EVAL [**germ_eval**]	4,546	45.93 %	33.76 %	53.70 %	19.94 %
German	HASOC [**10.1145/3368567.3368584**]	1,431	69.35 %	11.63 %	32.42 %	20.79 %
Japanese	Inspection AI [a]	324	25.86 %	35.93 %	16.98 %	-18.95 %
Japanese	LLM_JP [**DBLP:journals/corr/abs-2407-03963**]	1,662	10.02 %	44.72 %	45.43 %	0.71 %
Portuguese (Brazil)	OffCom [**brasnam**]	577	44.14 %	19.55 %	26.86 %	7.31 %
Portuguese (Brazil)	OLID [**10.1007/s10579-023-09657-0**]	5,534	20.40 %	85.39 %	94.00 %	8.62 %
Portuguese (Brazil)	ToLD [**leite-etal-2020-toxic**]	15,065	28.26 %	44.07 %	49.98 %	5.91 %
Russian	Abusive [**DBLP:conf/raslan/AndrusyakRK18**]	1,184	40.80 %	32.70 %	53.97 %	21.27 %
Russian	South_Park [**saitov-derczynski-2021-abusive**]	13,155	17.13 %	32.83 %	32.69 %	-0.14 %

[a]https://github.com/inspection-ai/japanese-toxic-dataset

Table 7: 单语（BERT, RoBERTa）和多语（XLM-RoBERTa）模型的比较。BERT-base 在游戏数据上取得了最佳的整体性能，而多语种模型保持了有竞争力的性能，并能够实现跨语言泛化到 MLSNT。

F1-Score	Model	GAME_1	GAME_2	Overall	MLSNT
Macro	bert-base	45.39 ± 1.01	40.94 ± 1.38	43.17	-
	roberta-base	42.85 ± 2.11	38.56 ± 2.69	40.71	-
	xlm-roberta-base	42.78 ± 1.75	38.64 ± 1.39	40.71	42.12 ± 1.68
	xlm-roberta-base-adapted	42.82 ± 1.42	39.09 ± 1.59	40.96	42.51 ± 2.71
Weighted	bert-base	82.15 ± 0.41	86.81 ± 0.44	84.48	-
	roberta-base	81.95 ± 0.47	86.67 ± 0.58	84.31	-
	xlm-roberta-base	81.70 ± 0.30	86.42 ± 0.29	84.06	86.44 ± 0.33
	xlm-roberta-base-adapted	81.79 ± 0.51	86.46 ± 0.32	84.13	86.15 ± 0.32

Table 9: 多语言人工标注数据集概览。

Language	Name	Task	Platform	Lines	Cost*
Chinese (Simplified)	COLD	Offensive	Zhihu, Weibo, etc. (SNS)	37,480	$ 3.64
Chinese (Simplified)	SWSR	Sexism	Weibo (SNS)	8,969	$ 0.89
Chinese (Simplified)	TOXICN	Toxicity & more	Zhihu, Tieba (SNS)	12,011	$ 1.15
Chinese (Traditional)	TOCAB	Profanity	PTT, a famous BBS site	104,002	$ 9.78
French	MLMA	Hate	Twitter	4,014	$ 0.38
German	GAHD	Hate	News, synthetic data	10,996	$ 1.04
German	GERM_EVAL	Offensive	Twitter	8,407	$ 0.81
German	HASOC	Offensive, Hate	Twitter and Facebook	4,669	$ 0.45
Japanese	Inspection AI	Toxicity	–	437	$ 0.04
Japanese	LLM_JP	Toxicity	– (long form text)	1,847	$ 0.51
Portuguese (Brazil)	OffCom	Offensive	Posts	1033	$ 0.10
Portuguese (Brazil)	OLID	Toxicity	Twitter, YouTube, and more	6,952	$ 0.66
Portuguese (Brazil)	ToLD	Offensive	Twitter	21,000	$ 1.99
Russian	Abusive	Abusive	Video comments	2,000	$ 0.20
Russian	South_Park	Hate Speech	South Park (video subtitles)	15,875	$ 1.57

```
You are a chat moderator for a game company. Your task is to classify whether the given line of game chat is toxic or not.

Below are the definitions for "toxic" categories.

    Definitions:
        1. "Threats (Life Threatening)" - concrete intention to cause harm, injury or other hostile actions to someone or to oneself
        2. "Minor Endangerment" -  intention or mentioning of where a child is / will be in a potentially harmful, dangerous, or
             inappropriate situation (e.g. grooming, neglect, pedophilia, child pornography, threats to a child, a child's personally
             identifiable information).
        3. "Threats (Non-Life Threatening)" - provocative statement that would cause a disruption in another player's life or game
             experience, without a clear intent to cause real-life injury or damage.
        4. "Hate" -  verbal abuse (e.g. intimidation, ridicule, derogatory or insulting remarks) based on another person's or group of
             people's actual or perceived identity (e.g., race, religion, color, sex, gender identity, national origin, age, disability,
             sexual orientation, genetic information).
        5. "Sexual Content / Harassment" - any text that contains, mentions or coerces other players with implicit / explicit sexual
             language, overtones, behavior or favors (e.g. euphemisms, flirting, innuendos, unwanted sexual advances, sexual requests)
        6. "Extremism" - political term, describes activities that are not in accordance with norms of the state; those who are fully
             intolerant toward others and reject the way of problem solving and existing social order.
        7. "Insults" - intent to damage the dignity of another person (or group of people), undermining their reputation or undermining
             their self-esteem.
        6. "Controversial / Potentially Toxic Topic" - Any text that contain any of the following subtopic(s):
            a. "Abortion": any mentions pro or against the termination of pregnancy
            b. "Religion": any text relating to / believing in a religion
            c. "Politics": any text related to how land is governed or how governments make rules and laws to manage human society.
            d. "Vulgar Content": any text that is ugly and crass (e.g. a crude joke)
            e. "Shocking / Disgusting Content": any text that contains gruesome, graphic or disgusting accounts / imagery
            f. "Hard Drugs": any mentions of strong and usually illegal drugs, likely to cause addiction
            g. "Alcohol": any mentions of alcohol or specific type of alcohol (e.g. beer, wine, vodka, ethanol, ethyl)
            h. "PII": stands for Personally Identifiable Information, any information that can be used to tie back to a person.
            i. "Trolling": deliberate intent of a player to disrupt another player's online experience.
            j. "Cheating": intent of using various methods / third party apps to create an advantage beyond normal gameplay
            k. "Scams and Advertisements": fraud, scamming (e.g. phishing), posting inappropriate links (malware, dangerous websites),
                 advertisement of websites, services, cheats or other products.
            l. "Spamming": excessive sharing of the same / similar words, phrases or emojis, repeatedly sharing of nonsensical comments
            m. "Competitors": mention of a person, product or company related to the game at hand.
            n. "Other Offensive Content": any text that is not covered by any other class / subtopic and is offensive and / or harms a
                 player's reasonable enjoyment of the game.

Use the following step-by-step instructions when classifying the line.

    Step 1 - The user will provide you with preceding chat lines as CONTEXT and text for the CURRENT_LINE.
    Step 2 - Familiarize yourself with the number of unique players and the lines they typed out.
    Step 3 - Based on the CONTEXT, decide if the CURRENT_LINE is "toxic" or "non-toxic". A line is toxic if at least one span belongs to one
           of the toxic categories. If "non-toxic", we skip step 4.
    Step 4 - Identify and classify the spans of toxic words based on the previously provided definitions. Below are some guidelines.
        a. A span can belong to one or more toxic category.
        b. If a span is classified as "Controversial / Potentially Toxic Topic", please also include the subtopic(s) the span is related to.
        c. "Threats (Life Threatening)" and "Threats (Non-Life Threatening)" are exclusive of each other.
        d. "Extremism" and ""Controversial / Potentially Toxic Topic", with the subtopic of "Politics" would appear together.

Think about it step by step.

Return in a json format where the first field is the "overall_category", being either "toxic" or "non-toxic".
If the overall_category is "toxic", return in the field "spans" a list which contains the corresponding fields {"text": "", "category: [""],
      }.
Example:
1. {"overall_category": "toxic"
 "spans": [
    {"text": "retard", "category": ["Insults"]}
 ]
}
2. {"overall_category": "non-toxic"}
```

Figure 2: 用于标签生成的系统提示（v1）。