STAND-Guard: A Small Task-Adaptive Content Moderation Model

Disclaimer: The paper may contain usage of potentially offensive, sensitive or mature content.

Minjia Wang*[1,2], Pingping Lin[1], Siqi Cai[1,3], Shengnan An[1,4], Shengjie Ma[1,5], Zeqi Lin[1], Congrui Huang[1], Bixiong Xu [1]

[1]Microsoft, [2]Harvard University, [3]Peking University, [4]Xi'an Jiaotong University, [5]University of Illinois Urbana-Champaign

[1] {v-minjiawang, pinlin, v-siqicai, t-shengnanan,v-shengjiema, Zeqi.Lin, conhua, bix}@microsoft.com [2]minjiawang@g.harvard.edu, [3]2201210579@pku.edu.cn, [4] an1006634493@stu.xjtu.edu.cn, [5]sm138@illinois.edu

Abstract

Content moderation, the process of reviewing and monitoring the safety of generated content, is important for development of welcoming online platforms and responsible large language models. Content moderation contains various tasks, each with its unique requirements tailored to specific scenarios. Therefore, it is crucial to develop a model that can be easily adapted to novel or customized content moderation tasks accurately without extensive model tuning. This paper presents STAND-GUARD, a Small Task-Adaptive coNtent moDeration model. The basic motivation is: by performing instruct tuning on various content moderation tasks, we can unleash the power of small language models (SLMs) on unseen (out-of-distribution) content moderation tasks. We also carefully study the effects of training tasks and model size on the efficacy of cross-task fine-tuning mechanism. Experiments demonstrate STAND-Guard is comparable to GPT-3.5-Turbo across over 40 public datasets, as well as proprietary datasets derived from real-world business scenarios. Remarkably, STAND-Guard achieved nearly equivalent results to GPT-4-Turbo on unseen English binary classification tasks. [1]

1 Introduction

Ensuring content safety is essential for online communities and social media platforms to maintain a friendly communication environment (Arora et al., 2023). With the rapid development of large language models (LLMs), content moderation has also become crucial for service providers to preserve model quality and safeguard user interactions (Markov et al., 2023).

Industries are developing automated content moderation algorithms to ensure online content safety and integrity. Recent advancements in deep learning have established supervised training of lightweight classifiers as a typical paradigm (Markov et al., 2023). This approach provides a low-cost and efficient way to filter undesired content. However, it also faces challenges, such as aligning sufficient training data with evolving community policies, and updating human reviewers on new harmful categories. Even with adequate training data, these classifiers, which are trained on fixed and labeled datasets for a specific task, may still struggle to cope with the diversity and complexity of textual content. They are inflexible to transfer to out-of-distribution tasks. On the other hand, while the success of generative LLMs like GPT-4 motivates their use in content moderation[2], this approach has limitations. When comparing the price of compute instances (V100 or A100 GPU) to the billing price of GPT models, it becomes evident that the cost of hosting LLMs is substantially high. Additionally, the risk of a single LLM's vulnerabilities being exploited by malicious actors presents significant challenges. These factors highlight the lack of practicality of these methods in real-world business scenarios.

Thus, we need to build a content moderation model which is much smaller and cheaper than those LLMs, like GPT-4, but still have enough domain knowledge and adaptability to handle new tasks with or without few-shot examples. Now this approach raises several questions: 1) How well can the moderation model cope with out-of-distribution data that may occur in real-world scenarios? 2) How can we obtain data to generate the model for content moderation, given that human annotations are costly and scarce? 3) How much data and how many tasks do we need to train the model effectively? Is there a trade-off between the number of tasks and the model's performance, or does more

* Work done during the internship at Microsoft.

[1]Previous presentation: https://arxiv.org/abs/2411.05214

[2]https://openai.com/index/using-gpt-4-for-content-moderation/

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 1–20
January 19–24, 2025.

data always lead to better results?

Nowadays, small language models (SLMs), like Mistral-7B (Jiang et al., 2023), Gemma (Team et al., 2024) and Phi-3 (Abdin et al., 2024) have shown impressive performance or even superiority in some domains. Additionally, hosting a SLM requires fewer resources. These inspire us to address the aforementioned challenges by fine-tuning these SLMs (Ma et al., 2023; Zhang et al., 2024; Umansky et al.). We propose a cross-task fine-tuning method specifically tailored for SLMs, focusing on content moderation domain. Our work contributes in three significant ways:

- We present a methodology, namely cross-task fine-tuning, to fine-tune SLMs for novel content moderation tasks, specifically for out-of-distribution data.
- We categorize public datasets into different tasks and use them for cross-task fine-tuning in content moderation. Through various experiments, we demonstrate the potential of cross-task fine-tuning a business model with public datasets, making it highly practical.
- We develop a unified task-adaptive model, STAND-Guard, through cross-task fine-tuning. We evaluate the model on both public and proprietary business datasets. The model surpasses GPT-3.5-Turbo on in-distribution data, and performs on par with GPT-3.5-Turbo on out-of-distribution tasks. Notably, STAND-Guard achieves comparable results to GPT-4-Turbo on unseen (out-of-distribution) English binary classification tasks.

2 Related Work

Advancements in Large and Small Language Models (LLMs and SLMs) have made them viable for various tasks, including content moderation. These models can be utilized through two main methods: prompting and fine-tuning.

Prompting involves providing the LLM/SLM with a specific query or instruction, which it then uses to generate a response. In terms of content moderation, the prompt generally incorporates the moderation guidelines, along with the content subject to review (Kolla et al., 2024). There are many prompting strategies (Guo et al., 2023; Franco et al., 2023; Kumar et al., 2024; Zhang et al., 2023, 2024), nevertheless, crafting prompts that accurately reflect the moderation guidelines while also enhancing the performance of language models demands significant human intervention and computational resources. Therefore, we do not primarily focus on prompt engineering in this study.

Fine-tuning, on the other hand, involves adjusting the LLM's or SLM's parameters to better suit a particular task such as content moderation. Some methods update all model parameters (Ghosh et al., 2024), which is a resource-intensive process due to the substantial size of language models. Consequently, more efficient alternative methods have been developed, which modify only a subset of parameters. Techniques under this category include adding task-specific layers (Wullach et al., 2021; Markov et al., 2023; Houlsby et al., 2019; Sen et al., 2024), LoRA (Hu et al., 2021; Ma et al., 2023) and prompt-tuning (Li and Liang, 2021; He et al., 2023; Liu et al., 2022; Markov et al., 2023; Qiao et al., 2024; Lester et al., 2021; Yuan et al., 2024). However, many of these approaches are restricted to specific content moderation tasks or undesired categories (Markov et al., 2023; Guo et al., 2023; Kolla et al., 2024), limiting their capability to generalize to new tasks and categories.

3 Methodology

3.1 Cross-Task Fine-Tuning

A *task* is an annotation process that determines if content requires modification, or identifies the types of harm or targeted groups involved. Each task is inherently linked with a guideline that outlines the procedure for the annotation process.

It is well-established that fine-tuning boosts the performance of *in-distribution tasks* (Ma et al., 2023), i.e., tasks encountered during fine-tuning. However, the computational intensity required to fine-tune a model for every task is considerable. Moreover, it is not feasible in actual business scenarios. The question then arises: how can fine-tuned models sustain or even enhance their performance when dealing with *out-of-distribution tasks*, i.e., tasks not present during fine-tuning?

To answer this question, we propose cross-task fine-tuning, which enhances the diversity of training tasks without increasing the total number of tasks or the number of samples used during fine-tuning.

3.2 Building the Training Set

Our primary goal is to design a training set that is both diverse and minimal, while still achieving a substantial increase in performance.

To this end, we first group content moderation tasks into categories and subcategories. Based on Wang et al. (2023b), we developed a two-level taxonomy for content moderation tasks, categorizing them into 4 primary categories and 8 subcategories. The 4 primary categories are *Malicious Actions*, *Discrimination / Exclusion / Toxicity / Hateful / Offensive*, *Information Hazards* and *Misinformation Harms*. Please refer to Appendix A for detailed definitions.

Then, we curate a compact training set that encompasses all the subcategories, utilizing only a single private dataset and two public datasets.

3.3 Fine-Tuning Models

Problem definition. Let's consider a content moderation task characterized by a guideline G, and a corresponding dataset represented as $\mathcal{D} = (G, \{x_i, y_i\}_{i=0}^{N-1})$. G is the moderation guideline in a human-readable format that describes the annotation standards, and specifies the output format of language models. x_i signifies the input content for the sample indexed at i, and y_i which falls in the set $\{0, 1, ..., K_{\mathcal{D}} - 1\}$ indicates the respective ground truth label. Given $\{G, x_i\}$, the goal for language models is to predict y_i.

Guideline generation. A guideline G comprised of two parts: 1) Definitions of the undesired content. For public datasets, these are extracted from the dataset description or the original paper if available; otherwise, they are generated by GPT-4-Turbo (see Appendix B.1 for details). For private datasets, our internal guidelines are used. 2) The classification process, which specifies the label set (e.g., binary classification or multi-class classification) , factors to consider during classification, and the expected output format.

Fine-tuning with QLoRA We chose to fine-tune SLMs based on QLoRA (Dettmers et al., 2024), which combines Quantization and Low-Rank Adapters to allow for efficient fine-tuning. The input during fine-tuning is the guideline G and the content x_i to be reviewed. An illustrative example of these input prompts can be found in Appendix B.2. The expected output is "Label: y_i" for each training sample. Note that we do not primarily focus on prompt engineering in this work, and just utilize the same prompt format across various tasks, baselines, and models.

4 Experimental Setup

4.1 Implementation Details

STAND-Guard uses Mistral-7B (Jiang et al., 2023) v0.1, a 7-billion-parameter model from Mistral AI, as backbone model. To assess the impact of the backbone model's size on the efficacy of cross-task fine-tuning, we compare the performance of STAND-Guard with models underpinned by different backbones: Phi-3-mini-128k-instruct (Abdin et al., 2024) (3.8 billion parameters) and Mixtral-8×7B (Jiang et al., 2024) v0.1 (47 billion parameters). They were fine-tuned, inferred and evaluated using the process and configuration detailed in Appendix C.

4.2 Baseline Models

We benchmark STAND-Guard against two sets of baseline models: task-specific models and general models. *Task-specific models* are trained for specific tasks but do not accommodate the input of custom policies, including Perspective API and OpenAI Content Moderation API. *General models* are designed to accept guideline as in-context input to steer the classification of the input text, including LlamaGuard, GPT-3.5-Turbo and GPT-4-Turbo. We provide a brief overview of these baselines in Appendix D.

4.3 Data Preparation

We have collected data from related research as well as various public repositories, with a summary provided in Appendix E. We have maintained the separation between the training and test sets for each dataset to ensure no overlap between them. The statistics of the training and evaluation datasets are presented in Tables 8 and 9.

To ensure a fair comparison, for each task, we will utilize the same prompt across baselines and models.

Training dataset. As mentioned in Section 3.2, to build a training set as minimal as possible, only the following three datasets are used as the training material.

PKU-Alignment BeaverTails and *PKU-Alignment SafeRLHF* (Ji et al., 2024) are datasets for safety alignment in LLMs including helpfulness and harmlessness. The datasets comprises dozens of tasks. We utilize its data solely for safety assessment purposes and convert each task into a binary classification task.

Private dataset is a collection curated from our business context, comprising texts that have been manually annotated for five distinct categories, including labels for sexual content, self-harm, violence, hate speech and jailbreak.

Evaluation dataset. We evaluate our methods against the two groups of models described in Section 4.2.

For *task-specific models*, which are designed to handle only specific types of content moderation tasks, we adopt the approach of Markov et al. (2023) to conduct comparisons across only 4 datasets primarily associated with hate speech, offensive language, and toxicity. These datasets include the *OpenAI Content Moderation dataset* (*OpenAI CM* for short) (Markov et al., 2023), *Jigsaw*[3], *TweetEval* (Barbieri et al., 2020), and *White Supremacist* (De Gibert et al., 2018) as shown in Appendices E and F.2. Due to the rate limits imposed by the Perspective API and OpenAI Content Moderation API, we sampled 5,000 entries from the entire Jigsaw dataset for our analysis.

For *general models*, which accommodate input based on custom guidelines, we conduct a comprehensive comparison across 42 datasets and 80 tasks.

5 Results and Analysis

5.1 In-Distribution Tasks

Table 1 presents the F1 scores on in-distribution tasks, from which we can draw three conclusions.

1) STAND-Guard, fine-tuned with cross-task learning, markedly surpasses the performance of the vanilla Mistral-7B, showcasing the effectiveness of fine-tuning on in-distribution tasks.

2) STAND-Guard outperforms GPT-4-Turbo, one of the most advanced LLMs, in content moderation. This indicates that specialized, fine-tuned smaller models can excel in specific tasks compared to a generic LLM. This insight implies that if training data is attainable in a business scenario, we can employ it for fine-tuning in order to achieve results comparable to those of GPT-4-Turbo. Appendix G includes an error analysis for GPT-4-Turbo.

3) The F1 score for the same model varies widely across datasets, despite the use of a uniform guideline generation method. This variation reflects the intrinsic differences between tasks, as detailed in a case study in Appendix I.

[3]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

5.2 Out-of-Distribution Tasks

5.2.1 Main Results

Table 2 presents the F1 scores for out-of-distribution English binary classification tasks across a broad spectrum of tasks.

STAND-Guard vs. vanilla models. Upon examing Table 2, a conclusion emerges: the results highlight the performance improvements of the model that underwent fine-tuning on selected tasks when assessed against a wide array of out-of-distribution tasks, in comparison to the vanilla model. It also shows that the fine-tuned model has achieved considerable gains in relatively novel tasks such as *irony detection*, *harassment detection*, and *toxicity detection*, surpassing the vanilla model's performance in these areas. These findings robustly endorse the efficacy of cross-task knowledge transfer, as it demonstrates the model's enhanced adaptability and generalization capabilities across various datasets and tasks.

STAND-Guard vs. GPT models. Additionally, the results from Table 2 indicate that the model, fine-tuned via cross-task methods, not only exceeds the performance of the larger GPT-3.5-Turbo ($0.528 \rightarrow 0.577$) but also exhibit small performance drop-off when compared to GPT-4-Turbo on binary classification tasks represented in English.

STAND-Guard vs. task-specific API models. Table 10 in Appendix H presents the results obtained using Perspective API and OpenAI Content Moderation API on datasets concerning hate and offensive language, following Markov et al. (2023)'s methodology. It demonstrates that STAND-Guard is the only task-adaptive model that outstrips task-specific API models. Moreover, STAND-Guard not only achieves the best results among all baselines—including API models trained on datasets related to hate speech and offensive language—but it also outshines the performance of GPT-3.5-Turbo and GPT-4-Turbo.

It is important to acknowledge that tasks derived from different datasets might show inconsistencies, which can be attributed to nuanced differences in their underlying concepts or definitions. We further explore the correlation between task semantic similarity and classification quality improvements on these tasks in Appendix L.

5.2.2 In-Context Learning Capability

We carried out further experiments to demonstrate that STAND-Guard, once fine-tuned via cross-task learning, retains the ability to perform in-

Dataset	Task	LlamaGuard	GPT-3.5-Turbo	GPT-4-Turbo	Mistral-7B	**STAND-Guard**
PKU-Alignment BeaverTails	Animal Abuse	0.580	0.341	0.694	0.438	0.742
	Child Abuse	0.553	0.176	0.372	0.325	0.815
	Controversial Topics, Politics	0.034	0.056	0.043	0.114	0.446
	Discrimination, Stereotype	0.618	0.348	0.330	0.456	0.731
	Drug Abuse, Weapons	0.611	0.457	0.317	0.419	0.746
	Financial & Property Crime	0.592	0.521	0.515	0.539	0.744
	Hateful & Offensive Language	0.529	0.328	0.326	0.254	0.670
	Misinformation	0.037	0.060	0.066	0.052	0.082
	Non-Violent Unethical Behavior	0.207	0.384	0.418	0.199	0.655
	Privacy Violation	0.303	0.177	0.449	0.288	0.800
	Self Harm	0.522	0.063	0.078	0.110	0.727
	Sexually Explicit	0.537	0.358	0.580	0.410	0.667
	Terrorism, Organized Crime	0.067	0.143	0.097	0.191	0.196
	Violence	0.253	0.672	0.681	0.397	0.800
PKU-Alignment Safe-RLHF	Unsafe	0.580	0.763	0.818	0.492	0.871
Private	Hate	0.700	0.745	0.697	0.642	0.827
	Self Harm	0.654	0.573	0.707	0.556	0.856
	Sexual	0.335	0.660	0.800	0.010	0.802
	Violence	0.324	0.538	0.719	0.486	0.745
AVG		0.423	0.388	0.458	0.336	0.680

Table 1: F1 scores on in-distribution tasks under zero-shot setting. Jailbreak in our private dataset does not have a test set.

Dataset	Task	LlamaGuard	GPT-3.5-Turbo	GPT-4-Turbo	Mistral-7B	**STAND-Guard**
CallMeSexist	Sexism	0.145	0.631	0.639	0.228	0.724
Civil-Comments	Insult	0.533	0.674	0.801	0.599	0.787
	Obscenity	0.028	0.347	0.346	0.609	0.179
	Severe Toxicity	0.481	0.416	0.141	0.582	0.530
	Sexually Explicit	0.016	0.142	0.029	0.483	0.134
	Threat	0.080	0.376	0.188	0.454	0.163
	Toxicity	0.494	0.730	0.474	0.573	0.759
Commonsense Morality	Ethics	0.014	0.809	0.874	0.711	0.735
CrowS-Pairs	Bias	0.675	0.707	0.778	1.000	1.000
DecodingTrust	Stereotype	0.985	0.943	1.000	1.000	1.000
DynaHate	Hate	0.150	0.827	0.834	0.612	0.673
Exaggerated Safety	Safety	0.020	0.882	0.950	0.038	0.829
HASOC (English)	Hate, Offensive	0.038	0.367	0.576	0.323	0.679
HateCheck	Hate	0.829	0.949	0.961	0.814	0.867
HateEval	Hate	0.666	0.005	0.655	0.639	0.462
HatemojiCheck	Hate	0.256	0.825	0.920	0.774	0.836
HateXplain	Hate	0.788	0.796	0.820	0.220	0.782
Jigsaw	Identity Hate	0.232	0.111	0.254	0.021	0.281
	Insult	0.501	0.261	0.447	0.080	0.416
	Obscene	0.167	0.322	0.534	0.084	0.512
	Severe Toxic	0.112	0.065	0.141	0.011	0.085
	Threat	0.221	0.030	0.253	0.006	0.300
	Toxic	0.549	0.359	0.474	0.127	0.551
OpenAI CM	Harassment	0.161	0.255	0.268	0.077	0.726
	Self Harm	0.000	0.292	0.630	0.099	0.928
Reddit Content Moderation	Rule Moderation	0.002	0.467	0.445	0.159	0.126
Scruples Anecdotes	Ethics	0.000	0.445	0.555	0.061	0.427
Social Bias Inference Corpus (SBIC)	Intentionally Offensive	0.707	0.726	0.722	0.665	0.709
	Potentially Offensive	0.738	0.738	0.731	0.633	0.728
	Sexually Offensive	0.509	0.666	0.641	0.689	0.195
SWAD	Swear	0.007	0.447	0.551	0.415	0.539
ToxiGen	Toxic	0.729	0.779	0.815	0.497	0.601
TrustworthyLLM	Safety	0.225	0.571	0.874	0.402	0.590
TweetEval	Hate	0.650	0.686	0.486	0.308	0.556
	Irony	0.061	0.685	0.780	0.005	0.685
	Offensive	0.381	0.582	0.525	0.573	0.682
USElectionHate	Hate	0.392	0.346	0.504	0.034	0.392
White Supremacist	Hate	0.503	0.796	0.711	0.582	0.739
AVG		0.343	0.528	0.588	0.400	0.577

Table 2: F1 scores on out-of-distribution tasks (binary classification, English data) under zero-shot setting.

context learning for new tasks. Utilizing the out-of-distribution datasets and tasks outlined in Section 4.3, we chose a subset of tasks for which training data was originally available (but not included in our training set) and applied Retrieval Augmented Generation (RAG) (An et al., 2023; Hu et al., 2022) to facilitate annotation for the SLM. Specifically, we dynamically selected the 10 few-shot samples most relevant to the content needing classification. Relevance was determined by calculating the cosine similarity between each training sample's embedding and the embedding of the content under review. These 10 samples were then appended after the guideline in order of ascending similarity.

Dataset	Task	STAND-Guard w/ RAG F1	Prec	Rec	STAND-Guard F1	Prec	Rec
PKU-Alignment-BeaverTails-Eval	Unsafe	0.583	0.665	0.593	0.513	0.695	0.526
Korean Hate Speech (Korean)	Hate	0.784	0.913	0.686	0.240	0.978	0.137
	Aggressiveness	0.941	0.944	0.941	0.382	0.484	0.450
	Bias	0.998	0.998	0.998	0.698	0.758	0.766

Table 3: F1 scores, precisions, recalls on out-of-distribution tasks under few-shot (w/ RAG) and zero-shot setting. The results show that fine-tuned task-adaptive model retains in-context learning ability.

Table 3 presents a comparative analysis of the performance of the same model, fine-tuned through cross-task learning, with and without the incorporation of RAG. The results confirm that STAND-Guard retains its in-context learning capabilities. Notably, when combined with RAG, STAND-Guard is capable of attaining competitive performance on par with the zero-shot capabilities of GPT-4-Turbo. Furthermore, the integration of RAG provides a tangible advantage for content moderation tasks within the fine-tuned SLM framework, enhancing both precision and recall metrics.

5.2.3 Multi-Lingual and Multi-Class Tasks

We also conduct an in-depth examination of performance across **multi-lingual** tasks and **multi-class classification** tasks. The results are displayed in Tables 11 and 12, respectively. Under such experimental settings, the overall performance of the fine-tuned model exhibits a reduction when compared to both GPT-3.5-Turbo and GPT-4-Turbo. This underlines the significance of maintaining a close alignment between the distribution of the training and testing data. This drop is due to fine-tuning with exclusively English binary classification data, highlighting the necessity for a more diverse training corpus to achieve optimal performance across varied linguistic contexts and task complexities.

5.3 Ablation Study

Table 4 presents the results of our training data ablation study. We systematically removed portions of the training data, initially detailed in Section 4.3, to evaluate cross-task knowledge transfer. Two additional experiments were conducted: 1) **STAND-Guard (w/o Private)**: excluded our proprietary datasets from the full training set. 2) **STAND-Guard (w/o Hate Offensive)**: removed data related to hate speech and offensive content from the full training set. From Table 4, we can draw three conclusions:

1) Even without proprietary datasets, STAND-Guard (w/o Private) showed significant improvements over vanilla Mistral-7B on private datasets. This finding is encouraging as it suggests that individuals can align a language model with business-specific guidelines by fine-tuning it solely on publicly available content moderation datasets, rather than relying on business data. This approach has the potential to save substantial time and resources that would otherwise be spent on collecting human-labeled data for the training set.

2) Despite the absence of hate speech and offensive content data, STAND-Guard (w/o Hate Offensive) still surpasses vanilla Mistral-7B. This indicates that tasks not directly related to hate or offensive language can still contribute positively to the detection of such content. This outcome further validates the efficacy of cross-task fine-tuning.

3) The model that was fine-tuned using the complete training dataset either outperformed or matched the performance of the two models trained on partial datasets across all tasks. This suggests that the private dataset (or the hate detection dataset) is capable of transferring knowledge to virtually all tasks, or at the very least, does not diminish the quality of detection.

5.4 Influence of Model Size

We analyze the impact of model size on cross-task fine-tuning by comparing the performance of three backbone models of various sizes: Phi-3-mini (3.8B parameters), STAND-Guard (7B parameters), and Mixtral-8×7B (47B parameters). Table 5 presents the average F1 scores for in-distribution and out-of-distribution tasks for each model. Detailed metric values for each task are provided in Appendix M. As shown, larger backbone models generally exhibit better generalizability when fine-tuned across various tasks. Although cross-task fine-tuning can be employed across a variety of backbone models, a balance between hosting/inference cost and inference quality should be taken into account for business scenarios.

6 Conclusions

In this study, we introduced a cross-task fine-tuning approach and demonstrated its efficacy using pub-

Dataset	Task	Mistral-7B	STAND-Guard (w/o Private)	STAND-Guard (w/o Hate Offensive)	STAND-Guard
PKU-Alignment BeaverTails	Animal Abuse	0.438	0.697	0.716	0.742
	Child Abuse	0.325	0.792	0.857	0.815
	Controversial Topics, Politics	0.114	0.372	0.450	0.446
	Discrimination, Stereotype	0.456	0.734	0.734	0.731
	Drug Abuse, Weapons	0.419	0.745	0.698	0.746
	Financial & Property Crime	0.539	0.742	0.751	0.744
	Hateful & Offensive Language	0.254	0.663	**0.571**	0.670
	Misinformation	0.052	0.000	0.051	0.082
	Non-Violent Unethical Behavior	0.199	0.633	0.657	0.655
	Privacy Violation	0.288	0.782	0.791	0.800
	Self Harm	0.110	0.710	0.667	0.727
	Sexually Explicit	0.410	0.634	0.611	0.667
	Terrorism, Organized Crime	0.191	0.089	0.163	0.196
	Violence	0.397	0.791	0.796	0.800
PKU-Alignment Safe-RLHF	Unsafe	0.492	0.846	0.844	0.871
Private	Hate	0.642	**0.700**	**0.729**	0.827
	Self Harm	0.556	**0.813**	0.839	0.856
	Sexual	0.010	**0.639**	0.671	0.802
	Violence	0.486	**0.743**	0.705	0.745
AVG		0.336	0.638	0.647	0.680

Table 4: Ablation study on training data. STAND-Guard (w/o Private): our proprietary datasets are excluded from the training set. STAND-Guard (w/o Hate Offensive): data related to hate speech and offensive content are removed. F1 scores on out-of-distribution tasks (i.e., tasks not in the training set) are in bold.

Dataset/Task	Phi-3-mini (CT-FT)	STAND-Guard	Mixtral-8x7B (CT-FT)
in-distribution	0.581	0.680	0.671
out-of-distribution	0.488	0.533	0.577

Table 5: Average F1 scores for cross-task fine-tuned models of various sizes.

licly available content moderation datasets. Our findings reveal that fine-tuning a SLM exclusively with public content moderation data can yield robust performance in bespoke scenarios governed by custom guidelines. Furthermore, our approach enables knowledge transfer across tasks, even when the tasks are not closely related. By employing cross-task fine-tuning, we successfully developed a high-quality model that is comparable to GPT-3.5-Turbo on various tasks, and achieves nearly equivalent results to GPT-4-Turbo on brand new English binary classification tasks. This underscores the potential of our method as a competitive alternative in the realm of advanced language models.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? *arXiv preprint arXiv:2305.04835*.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Vinzent Seiler, Dennis M. Riehle, and Heike Trautmann. 2021. Rp-mod&rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *NeurIPS Datasets and Benchmarks*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Rogers P. de Pelle and Viviane P. Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2023. Analyzing the use of large language models for content moderation with chatgpt examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 1–8.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2023. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 61–61. IEEE Computer Society.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch Critch, Jerry Li Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. In *International Conference on Learning Representations*.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021b. What would jiminy cricket do? towards agents that behave morally. *arXiv preprint arXiv:2110.13136*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. 2021. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *arXiv preprint arXiv:2309.14517*.

Joao A Leite, Diego F Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv e-prints*.

Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *arXiv preprint arXiv:2310.03400*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. Beep! korean corpus of online news comments for toxic speech detection. *arXiv preprint arXiv:2005.12503*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6237–6246.

Wei Qiao, Tushar Dogra, Otilia Stretcu, Yu-Han Lyu, Tiantian Fang, Dongjin Kwon, Chun-Ta Lu, Enming Luo, Yuan Wang, Chih-Chun Chia, et al. 2024. Scaling up llm reviews for google ads content moderation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1174–1175.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 573–584.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Tanmay Sen, Ansuman Das, and Mrinmay Sen. 2024. Hatetinyllm: Hate speech detection using tiny large language models. *arXiv preprint arXiv:2405.01577*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Natalia Umansky, Maël Kubli, Karsten Donnay, Fabrizio Gilardi, Dominik Hangartner, Ana Kotarcic, Laura Bronner, Selina Kurer, and Philip Grech. Enhancing hate speech detection with fine-tuned large language models requires high-quality data.

USElectionHate. *https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/stance-hof/*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023b. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705.

Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. 2024. Rigorllm: Resilient guardrails for large language models against undesired content. *arXiv preprint arXiv:2403.13031*.

Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, and Konstantinos Psounis. 2024. Efficient toxic content detection by bootstrapping and distilling large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21779–21787.

Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Interpretable unified language checking. *arXiv preprint arXiv:2304.03728*.

A Categories of Content Moderation Tasks

Based on Wang et al. (2023b), we classified content moderation tasks into the following categories:

Malicious actions. This category encompasses tasks involve the modification of content that promotes or aids actions with potential harmful consequences. It can be divided into two subcategories: 1) Illegal Activities, which consist of content endorsing violence, threats, substance abuse, and more. 2) Unethical or Unsafe Actions, which cover content that encourages unhealthy practices, provides guidance for unsafe behaviors, or promotes harassment. Additionally, the text specifies that content pertaining to "Jailbreak" falls within this category, which includes attempts to circumvent safeguards and elicit unauthorized outputs from large language models.

Discrimination, exclusion, toxicity, hateful, offensive. This category involves tasks for addressing harmful and toxic online content. There are three subcategories: 1) Social Stereotype and Unfair Discrimination, which involves content that propagates prejudices or stereotypes against specific groups of people. 2) Toxic Language (Hate Speech), including toxic or offensive language. 3) Adult Content, which consists of explicit sexual material and graphic violence.

Information hazards. Tasks which aims to identifying content that presents potential information risks falls under this category. For instance, it targets content that probes for personal, private details such as sensitive financial data or personal contact details, which could potentially lead to serious repercussions.

Misinformation harms. Tasks for detecting content that disseminates misleading information or causes potential harm belongs to this category. It consists of two subcategories: 1) Disseminating False Information, such as spam, rumors and propaganda. 2) Causing Material Harm through Misinformation, which involves content that shares unverified or uncorroborated information, potentially leading to medical or financial harm.

B Prompt Generation for Public Content Moderation Dataset

B.1 Definition Generation

As outlined in Section 3.3, a guideline involves two components: the definitions of undesired content and the evaluation process. When the definitions are not readily available in the source dataset, we employ GPT-4-Turbo to generate these definitions.

This is accomplished through a few-shot learning process. For each sample, the prompt is the name of the undesired category (for instance, hate speech or violence), and the generated output is the definition of that category. We utilize the definitions from our private dataset as few-shot samples. The configuration of GPT-4-Turbo follows the setting mentioned in Appendix D.

B.2 Example Prompts

The text below serves as a sample prompt for the *Anatomy of Online Hate* dataset(Salminen et al., 2018), aimed at classifying the targets of hate speech. The first part describes the task and provides definitions for each target as outlined in Salminen et al. (2018). The second part, starting with "Key Points to Consider", clarifies the evaluation process and the data format. The third part is the content to be reviewed, namely "<content>".

Hate Speech Target Classification Guidelines

Your goal is to classify content based on its target, according to the 13 targets defined below. Only one target should be labeled for each content.

Definitions of Targets

- **None (Label: 0).** The content is not hateful, and therefore there is no target.
- **Financial Powers (Label: 1).** Hatred toward wealthy people and companies and their privileges. Pointing out their intentions to manipulate and commit crimes.
- **Political Issues (Label: 2).** Hate toward government, political parties and movements, war, terrorism, the flaws of the system.
- ...

Key Points to Consider

1. **Understanding the Categories:** Familiarize yourself with the definitions of the 13 harm categories. Understand what each category represents and the kind of content it covers.

2. **Single Category Labeling:** Even if a content can be associated with multiple categories, only one category should be labeled for each content. Choose the most prominent or harmful category.

3. **Input Format:** The content to be classified is located after "Content". This is the specific content that you should focus on for classification purposes.

4. **Output Format:** Begin your classification response with the term "Label: ", followed by the index number of the category. Index numbering is indicated in the parentheses and starts from 0. For instance, if the content is classified under the category "None (Label: 0)", your response should be "Label: 0".

Content

<content>

C Implementation Details

Fine-tuning. The fine-tuning process was conducted for 1 epoch, employing a batch size of 96. The learning rate is 1e-4 with a warm-up ratio of 0.03. For QLoRA, we set the rank to 64 and the scaling factor to 16. The dropout probability for LoRA is 0.05.

We chose Mistral-7B (Jiang et al., 2023) v0.1, a 7-billion-parameter language model that has been open-sourced by Mistral AI, as the backbone model for STAND-Guard and evaluate its effectiveness compared to the vanilla model and some other baselines. We further evaluate the influence of the backbone model's size by comparing models cross-task fine-tuned on Phi-3-mini-128k-instruct, which has 3.8 billion parameters (Abdin et al., 2024), and Mixtral-8×7 version 0.1 (Jiang et al., 2024) (47 billion parameters), with our STAND-Guard model. All the models adhere to identical training protocols mentioned above.

Inference. During inference, we assign a top_p value of 1.0, a temperature of 0.0 and a max_tokens of 100 for all the models.

Metrics. The **F1 score** is used as the evaluation metrics. For multi-class classification, we calculate the F1 metrics for each label, and find their average weighted by support (the number of true instances for each label)[4].

It should be noted that we classify any predictions that do not adhere to the schema outlined in the guideline as incorrect. Consequently, the F1 score are calculated based on the entire set of cases, rather than solely on those successfully parsed. Higher F1 values indicate better performance.

D Baseline Models

D.1 Task-Specific Models

These models are classifiers that are trained for specific tasks but do not accommodate the input of custom policies.

Perspective API. Perspective API [5] offers services for the detection of toxic and hateful content. It encompasses a range of categories, such as toxicity, severe toxicity, insult, profanity, identity attacks, threats, and sexually explicit material. For the purpose of comparison, we convert the scores returned by the API into binary outcomes using a threshold of 0.5.

OpenAI Content Moderation API. OpenAI Content Moderation API (Markov et al., 2023) is trained to detect a set of categories of undesired content, including sexual content, hateful content, violence, self-harm, and harassment. Similar to the settings for Perspective API, we binarize the scores provided by the API with a threshold value of 0.5.

D.2 General Models

These models refer to LLMs and SLMs that are designed to accept the guideline as an in-context input to steer the classification of the input text.

LlamaGuard. LlamaGuard (Inan et al., 2023) is fine-tuned for content moderation based on Llama2-7B. The first token of the output is adjusted to indicate if the content is "safe" or "unsafe", and the second token indicates the specific harmful category. We made slight modifications to the prompt's output schema for LlamaGuard to ensure compatibility with its pre-trained counterparts.

GPT-3.5-Turbo and GPT-4-Turbo. GPT-4 is considered to be the most powerful LLM to date, and GPT-4-Turbo [6] is a new version that supports longer context. There is ongoing work to integrate GPT models for content moderation[7]. For both GPT-3.5-Turbo [8] and GPT-4-Turbo, in addition to the common configuration shared by all models, we configured frequency_penalty and presence_penalty to 0.

E Task Analysis

Content moderation tasks are classified based on categories and subcategories described in Section 3.2. Table 6 and 7 show a detailed analysis of all the tasks used in this study. A mark (✓) denotes that the undesired content overlaps with the subcategory. It is noteworthy that even though some tasks share the same name, the definition of the undesired content can be different, highlighting the importance of developing a model that quickly adapts to diverse content moderation tasks.

F Data Statistics

F.1 Training Data

The statistics of training data is presented in Table 8. We conducted strategic sampling to guarantee that each task is represented in roughly equal proportions within the full training dataset.

F.2 Testing Data

Table 9 shows the statistics of tasks in the test set.

G Error Analysis of GPT-4

The performance difference on in-distribution tasks between zero-shot GPT-4-Turbo and STAND-guard is considerable (0.68 vs 0.46). Given that GPT-4-Turbo possesses outstanding zero-shot capability, it is necessary to delve into the errors of GPT-4-Turbo.

A common error in the PKU-Alignment Safe-RLHF dataset is role misinterpretation. This dataset features conversations between a user and a bot, with only the bot's response being subject to modification. GPT-4-Turbo often misclassifies based on the user's input rather than the bot's response. In contrast, STAND-Guard, having encountered such conversations during training, accurately identifies and flags inappropriate bot responses.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[5] https://www.perspectiveapi.com/

[6] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[7] https://openai.com/index/using-gpt-4-for-content-moderation/

[8] https://platform.openai.com/docs/models/gpt-3-5-turbo

Dataset	Task	Malicious Actions		Discrimination, Exclusion, Toxicity, Hateful, Offensive			Information Hazards	Misinformation Harms	
		Illegal Activities	Unethical or Unsafe Actions	Social Stereotype and Unfair Discrimination	Toxic Language (Hate Speech)	Adult Content	Private Information	Disseminate False or Misleading Information	Causing Material Harm by Disseminating Unverified Information
Anatomy of Online Hate	Hate				✓				
BIG-bench (German)	Gender Inclusive			✓					
BIG-bench (Hinglish)	Toxic				✓				
CallMeSexist	Sexism			✓					
Civil-Comments	Insult				✓				
	Obscenity				✓				
	Severe Toxicity				✓				
	Sexually Explicit					✓			
	Threat	✓							
	Toxicity				✓				
Commonsense Morality	Ethics		✓						
COVID-HATE	Hate				✓			✓	
CrowS-Pairs	Bias			✓					
DecodingTrust	Stereotype			✓					
DynaHate	Hate				✓				
Exaggerated Safety	Safety	✓	✓	✓			✓		
GermEval (German)	Offensive				✓				
HASOC (English)	Hate, Offensive				✓				
HASOC (German)	Hate, Offensive				✓				
Hate Speech and Offensive Language	Hate, Offensive				✓				
Hate Speech towards Foreigners (German)	Hate				✓				
HateCheck	Hate				✓				
HateEval	Hate				✓				
HatemojiCheck	Hate				✓				
HateXplain	Hate	✓			✓				
Jigsaw	Identity Hate			✓	✓				
	Insult				✓				
	Obscene				✓				
	Severe Toxic				✓				
	Threat	✓							
	Toxic				✓				
Jiminy Cricket	Ethics		✓						
Korean Hate Speech (Korean)	Hate			✓	✓				
	Aggressiveness				✓				
	Bias			✓					
OffComBR3 (Portuguese)	Offensive			✓	✓				
OpenAI CM	Harassment		✓						
	Hateful	✓			✓				
	Self Harm		✓						
	Sexual					✓			
	Violence	✓				✓			

Table 6: Task analysis based on the task taxonomy. A mark (✓) denotes that the definition of the undesired content overlaps with the subcategory of the content moderation task.

Dataset	Task	Malicious Actions: Illegal Activities	Malicious Actions: Unethical or Unsafe Actions	Discrimination, Exclusion, Toxicity, Hateful, Offensive: Social Stereotype and Unfair Discrimination	Discrimination, Exclusion, Toxicity, Hateful, Offensive: Toxic Language (Hate Speech)	Discrimination, Exclusion, Toxicity, Hateful, Offensive: Adult Content	Information Hazards: Private Information	Misinformation Harms: Disseminate False or Misleading Information	Misinformation Harms: Causing Material Harm by Disseminating Unverified Information
PKU-Alignment Safe-RLHF	Unsafe	✓	✓	✓	✓	✓	✓	✓	✓
PKU-Alignment BeaverTails	Animal Abuse	✓							
	Child Abuse	✓							
	Controversial Topics, Politics							✓	
	Discrimination, Stereotype			✓					
	Drug Abuse, Weapons	✓							
	Financial & Property Crime	✓							
	Hateful & Offensive Language				✓				
	Misinformation	✓	✓					✓	✓
	Non-Violent Unethical Behavior		✓						
	Privacy Violation						✓		
	Self Harm		✓						
	Sexually Explicit					✓			
	Terrorism, Organized Crime	✓							
	Violence	✓							
PKU-Alignment-BeaverTails-Eval	Unsafe	✓	✓	✓	✓	✓	✓	✓	✓
Private	Hate				✓				
	Self Harm		✓						
	Sexual					✓			
	Violence	✓							
	Jailbreak		✓						
Reddit Content Moderation	Rule Moderation	✓	✓	✓	✓	✓	✓	✓	
RP-Mod & RP-Crowd (German)	Offensive	✓		✓	✓			✓	
Scruples Anecdotes	Ethics		✓						
Social Bias Inference Corpus (SBIC)	Intentionally Offensive				✓				
	Potentially Offensive				✓				
	Sexually Offensive				✓				
SWAD	Swear				✓				
SWSR (Chinese)	Sexism			✓					
ToLD-BR (Portuguese)	Offensive			✓	✓				
	Homophobia			✓					
	Misogyny			✓					
	Racism			✓					
ToxiGen	Toxic				✓				
TrustworthyLLM	Safety	✓	✓			✓	✓		
TweetEval	Hate				✓				
	Irony		✓						
	Offensive				✓				
USElectionHate	Hate	✓		✓	✓				
White Supremacist	Hate			✓	✓				

Table 7: Task analysis based on the task taxonomy (continued). A mark (✓) denotes that the definition of the undesired content overlaps with the subcategory of the content moderation task.

Dataset	#Task	Task	Binary/Multiple Class(es)	#Data	#Data (%)	Harmful Ratio
PKU-Alignment BeaverTails	14	Animal Abuse	binary	10,000	5.7%	1.3%
		Child Abuse	binary	9,949	5.7%	0.7%
		Controversial Topics, Politics	binary	9,981	5.7%	3.5%
		Discrimination, Stereotype	binary	9,984	5.7%	8.8%
		Drug Abuse, Weapons	binary	10,000	5.7%	5.6%
		Financial & Property Crime	binary	9,942	5.7%	9.2%
		Hateful & Offensive Language	binary	10,000	5.7%	9.2%
		Misinformation	binary	10,000	5.7%	2.3%
		Non-Violent Unethical Behavior	binary	9,964	5.7%	17.4%
		Privacy Violation	binary	9,981	5.7%	5.4%
		Self Harm	binary	9,943	5.7%	0.8%
		Sexually Explicit	binary	9,941	5.7%	2.5%
		Terrorism, Organized Crime	binary	10,000	5.7%	1.1%
		Violence	binary	9,976	5.7%	25.5%
PKU-Alignment SafeRLHF	1	Unsafe	binary	10,000	5.7%	43.0%
Private	5	Hate	binary	3,733	2.1%	19.1%
		Self harm	binary	3,739	2.1%	11.9%
		Sexual	binary	3,794	2.2%	17.9%
		Violence	binary	3,734	2.1%	18.1%
		Jailbreak	binary	10,000	5.7%	10.4%

Table 8: Statistics of training data. #Data and #Data (%) indicate the number and proportion of training samples for the task, respectively. The Harmful Ratio indicates the proportion of entries with positive labels for the task.

Another error is target generalization, which occurs when guidelines focus on a specific target (e.g., child abuse), but the content is more general (e.g., abuse). GPT-4-Turbo tends to incorrectly flag such content. However, STAND-Guard, due to its fine-tuning process, is better equipped to handle these nuances.

H Comparison with Task-Specific Baselines

Table 10 contains the results obtained using task-specific API models on datasets concerning hate speech and offensive language, in accordance with the methodology outlined by Markov et al. (2023).

I Case Study

Cross-task fine-tuning enables the model to understand derogatory implications. The significant classification gain on the dataset *CallMeSexist*, which bears scant resemblance to the training set, serves as a case in point. Consider the following statement from *CallMeSexist*: *For real. Asking a woman to move smh!! naw no go"MENTION3988: MENTION3549 lmao IDC IDC! Call your male friends."* It is not explicitly demeaning towards women. Instead, it subtly suggests that a woman's presence is less important or valued than a man's. STAND-Guard succeeds in detecting implicit harmful intent in the statement.

Without external knowledge, all models struggle to flag the undesired content. For instance, when detecting controversial or political topics in *PKU-Alignment BeaverTails*, even GPT-4-Turbo achieves a F1 of only 0.043. This could be attributed to the intrinsic nature of the task. The conversations in this task encompass political figures, policies and current affairs, such as *Josef Mengele is most known for his horrific human experiments at Auschwitz during WWII, ...* To make correct predictions, knowledge about these topics are necessary, posing challenges for pre-trained models.

J Multilingual Tasks

Tables 11 shows the experimental results on non-English, out-of-distribution tasks under zero-shot setting.

K Multi-Class Classification Tasks

Tables 12 shows the experimental results on multi-class classification, out-of-distribution tasks under zero-shot setting.

L Visualization of Task Semantic Similarity and Classification Gains

We further explore the correlation between task semantic similarity and classification quality improvements on English binary classification tasks in greater depth in Figure 1. We obtain these semantic similarities by representing each task as a binary vector, as per the data presented in Appendix E. Following this, we calculate the cosine similarity between these tasks to determine their relative similarity. The analysis indicates that the fine-tuned model realized significant enhancements in classification quality over the vanilla model for tasks that closely resemble the training data, including *Exaggerated Safety*, *Jigsaw - Threat*, *TweetEval - Irony*, and *USElectionHate*. Intriguingly, we also observed benefits from cross-task knowledge

Dataset	#Task	Task	Binary/Multiple Class(es)	#Data	Harmful Ratio
PKU-Alignment BeaverTails	14	Animal Abuse	binary	3,021	1.5%
		Child Abuse	binary	3,021	0.9%
		Controversial Topics, Politics	binary	3,021	3.1%
		Discrimination, Stereotype	binary	3,021	9.8%
		Drug Abuse, Weapons	binary	3,021	5.0%
		Financial & Property Crime	binary	3,021	8.7%
		Hateful & Offensive Language	binary	3,021	10.0%
		Misinformation	binary	3,021	2.5%
		Non-Violent Unethical Behavior	binary	3,021	20.1%
		Privacy Violation	binary	3,021	5.1%
		Self Harm	binary	3,021	0.6%
		Sexually Explicit	binary	3,021	3.4%
		Terrorism, Organized Crime	binary	3,021	1.4%
		Violence	binary	3,021	24.0%
PKU-Alignment Safe-RLHF	1	Unsafe	binary	66,088	53.5%
Private	4	Hate	binary	1,000	30.3%
		Self harm	binary	1,000	19.7%
		Sexual	binary	1,000	19.2%
		Violence	binary	1,000	26.3%
OpenAI CM	5	Harassment	binary	1,444	5.3%
		Hateful	multiple	1,680	23.9%
		Self Harm	binary	1,447	3.5%
		Sexual	multiple	1,680	25.7%
		Violence	multiple	1,680	6.2%
TweetEval	3	Hate	binary	2,970	42.2%
		Irony	binary	784	39.7%
		Offensive	binary	860	27.9%
Jigsaw	5	Identity Hate	binary	63,978	1.1%
		Toxic	binary	63,978	9.5%
		Threat	binary	63,978	0.3%
		Insult	binary	63,978	5.4%
		Obscenity	binary	63,978	5.8%
		Severe Toxicity	binary	63,978	0.6%
White Supremacist (de Gibert et al., 2018)	1	Hate	binary	478	50.0%
Anatomy of Online Hate (Salminen et al., 2018)	1	Hate	multiple	3,222	73.4%
BIG-bench (German) (bench authors, 2023)	1	Gender Inclusive	binary	489	40.9%
CallMeSexist (Samory et al., 2021)	1	Sexism	binary	13,631	13.3%
Civil-Comments (Borkan et al., 2019)	6	Insult	binary	2,997	49.9%
		Obscenity	binary	1,998	49.9%
		Severe Toxicity	binary	2,985	49.9%
		Sexually Explicit	binary	1,990	50.2%
		Threat	binary	1,996	50.1%
		Toxicity	binary	2,997	49.9%
Commonsense Morality (Hendrycks et al., 2021a)	1	Ethics	binary	3,885	46.7%
COVID-HATE (He et al., 2021)	1	Hate	multiple	2,290	41.3%
CrowS-Pairs (Nangia et al., 2020)	1	Bias	binary	1,508	100.0%
DecodingTrust (Wang et al., 2023a)	1	Stereotype	binary	1,152	100.0%
DynaHate (Vidgen et al., 2021)	1	Hate	binary	41,255	54.0%
Exaggerated Safety (Röttger et al., 2023)	1	Safety	binary	450	44.4%
GermEval (German) bin (Wiegand et al., 2018)	1	Offensive	binary	3,532	34.0%
GermEval (German) multi (Wiegand et al., 2018)	1	Offensive	multiple	3,532	89.2%
HASOC (English) (Mandl et al., 2019)	1	Hate, Offensive	binary	1,153	25.0%
HASOC (German) (Mandl et al., 2019)	1	Hate, Offensive	binary	3,819	10.7%
Hate Speech and Offensive Language (Davidson et al., 2017)	1	Hate, Offensive	multiple	24,783	94.2%
Hate Speech towards Foreigners (German) (Bretschneider and Peters, 2017)	1	Hate	multiple	666	100.0%
HateCheck (Röttger et al., 2020)	1	Hate	binary	3,901	68.2%
HateEval (Basile et al., 2019)	1	Hate	binary	4,571	41.8%
HatemojiCheck (Kirk et al., 2021)	1	Hate	binary	3,930	67.5%
HateXplain (Mathew et al., 2021)	1	Hate	binary	1,924	59.4%
Jiminy-Cricket (Hendrycks et al., 2021b)	1	Ethics	multiple	3,986	50.4%
Korean Hate Speech (Korean) (Moon et al., 2020)	3	Hate	binary	471	68.4%
		Aggressiveness	multiple	471	66.0%
		Bias	multiple	471	27.4%
OffComBR3 (Portuguese) (de Pelle and Moreira, 2017)	1	Offensive	binary	1,250	33.5%
PKU-Alignment-BeaverTails-Eval (Ji et al., 2024)	1	Unsafe	multiple	700	92.9%
Reddit Content Moderation (Kumar et al., 2024)	1	Rule Moderation	binary	96,544	50.2%
RP-Mod & RP-Crowd (German) (Assenmacher et al., 2021)	1	Offensive	binary	57,410	50.0%
Scruples Anecdotes (Lourie et al., 2020)	1	Ethics	binary	6,159	25.0%
Social Bias Inference Corpus (SBIC) (Sap et al., 2019)	3	Intentionally Offensive	binary	3,462	50.0%
		Potentially Offensive	binary	5,892	50.0%
		Sexually Offensive	binary	3,462	50.0%
SWAD (Pamungkas et al., 2020)	1	Swear	binary	2,577	32.7%
SWSR (Chinese) bin (Jiang et al., 2022)	1	Sexism	binary	8,969	34.5%
SWSR (Chinese) multi (Jiang et al., 2022)	1	Sexism	multiple	8,969	34.5%
ToLD-BR (Portuguese) (Leite et al., 2020)	4	Offensive	binary	21,000	44.1%
		Homophobia	binary	21,000	1.6%
		Misogyny	binary	21,000	2.2%
		Racism	binary	21,000	0.7%
ToxiGen (Hartvigsen et al., 2022)	1	Toxic	binary	940	43.2%
TrustworthyLLM (Liu et al., 2023)	1	Safety	binary	5,904	14.0%
USElectionHate(USElectionHate)	1	Hate	binary	600	9.8%

Table 9: Statistics of test data. #Data and #Data (%) indicate the number and proportion of training samples for the task, respectively. For binary classification tasks, the Harmful Ratio indicates the proportion of entries with positive labels for the task. For multi-class tasks, the Harmful Ratio takes into account all instances that are marked with positive labels.

Dataset	Task	Perspective	OpenAI CM	LlamaGuard	GPT-3.5-Turbo	GPT-4-Turbo	Mistral-7B	**STAND-Guard**
Jigsaw (sampled)	Identity Hate	0.278	0.579	0.214	0.098	0.236	0.018	0.255
	Insult	0.482	0.498	0.469	0.233	0.425	0.077	0.410
	Obscene	0.531	0.225	0.161	0.304	0.545	0.081	0.508
	Threat	0.159	0.359	0.243	0.041	0.281	0.004	0.302
	Toxic	0.119	0.584	0.529	0.364	0.451	0.127	0.558
OpenAI CM	Harassment	0.290	0.327	0.161	0.255	0.268	0.077	0.726
	Hateful	0.716	0.732	0.729	0.669	0.671	0.577	0.732
	Self Harm	-	0.891	0.000	0.292	0.630	0.099	0.928
	Sexual	0.655	0.755	0.742	0.696	0.742	0.461	0.475
	Violence	0.922	0.949	0.927	0.815	0.707	0.674	0.671
TweetEval	Hate	0.249	0.295	0.650	0.686	0.486	0.308	0.556
	Irony	-	-	0.061	0.685	0.780	0.005	0.685
	Offensive	0.614	0.480	0.381	0.582	0.525	0.573	0.682
White Supremacist	Hate	0.584	0.508	0.503	0.796	0.711	0.582	0.739
AVG		0.467	0.552	0.412	0.465	0.533	0.262	0.588

Table 10: F1 scores on out-of-distribution tasks related to hate speech and offensive language detection under zero-shot setting. Task-specific baselines (Perspective and OpenAI CM) are included in the comparison.

Dataset	Task	LlamaGuard	GPT-3.5-Turbo	GPT-4-Turbo	Mistral-7B	**STAND-Guard**
BIG-bench (German)	Gender Inclusive	0.000	0.855	0.851	0.000	0.708
GermEval (German) bin	Offensive	0.012	0.679	0.670	0.574	0.501
GermEval (German) multi	Offensive	0.060	0.761	0.734	0.190	0.101
HASOC (German)	Hate, Offensive	0.038	0.219	0.302	0.217	0.417
Hate Speech towards Foreigners (German)	Hate	0.279	0.566	0.674	0.532	0.495
Korean Hate Speech (Korean)	Hate	0.206	0.712	0.824	0.012	0.240
	Aggressiveness	0.300	0.550	0.705	0.539	0.382
	Bias	0.736	0.706	0.739	0.458	0.698
OffComBR3 (Portuguese)	Offensive	0.340	0.640	0.722	0.424	0.558
RP-Mod & RP-Crowd (German)	Offensive	0.467	0.301	0.358	0.496	0.210
SWSR (Chinese) bin	Sexism	0.035	0.602	0.568	0.211	0.504
SWSR (Chinese) multi	Sexism	0.567	0.576	0.625	0.182	0.528
ToLD-BR (Portuguese)	Offensive	0.459	0.656	0.693	0.482	0.302
	Homophobia	0.093	0.130	0.428	0.034	0.140
	Misogyny	0.004	0.153	0.297	0.023	0.624
	Racism	0.062	0.055	0.258	0.015	0.215
AVG		0.229	0.510	0.591	0.274	0.414

Table 11: F1 scores on out-of-distribution tasks (non-English data) under zero-shot setting.

Dataset	Task	LlamaGuard	GPT-3.5-Turbo	GPT-4-Turbo	Mistral-7B	**STAND-Guard**
Anatomy of Online Hate	Hate	0.255	0.278	0.541	0.183	0.269
COVID-HATE	Hate	0.434	0.085	0.796	0.503	0.695
GermEval (German) multi	Offensive	0.060	0.761	0.734	0.190	0.101
Hate Speech and Offensive Language	Hate, Offensive	0.497	0.799	0.870	0.655	0.585
Hate Speech towards Foreigners (German)	Hate	0.279	0.566	0.674	0.532	0.495
Jiminy-Cricket	Ethics	0.329	0.633	0.646	0.407	0.598
Korean Hate Speech (Korean)	Aggressiveness	0.300	0.550	0.705	0.539	0.382
	Bias	0.736	0.706	0.739	0.458	0.698
OpenAI CM	Hateful	0.729	0.669	0.671	0.577	0.732
	Sexual	0.742	0.696	0.742	0.461	0.475
	Violence	0.927	0.815	0.707	0.674	0.671
PKU-Alignment-BeaverTails-Eval	Unsafe	0.150	0.467	0.448	0.136	0.513
SWSR (Chinese) multi	Sexism	0.567	0.576	0.625	0.182	0.528
AVG		0.462	0.585	0.684	0.423	0.519

Table 12: F1 scores on out-of-distribution tasks (multi-class classification) under zero-shot setting.

transfer on test tasks that deviate from the training tasks in similarity, such as *CallMeSexist* and *Jigsaw*. This suggests that the fine-tuning process imparts a degree of generalizability to the model, allowing it to effectively adapt and perform well even on tasks that are not directly semantically aligned with the original training data.

M Influence of Model Size

Table 13 and Table 14 show detailed F1 scores for cross-task fine-tuned models of various sizes.

Dataset	Task	Rel. F1 Gain	Avg. Sim.
CallMeSexist	Sexism	217.5%	0.0677
Civil-Comments	Insult	31.4%	0.1177
	Obscenity	-70.6%	0.1177
	Severe Toxicity	-8.9%	0.1177
	Sexually Explicit	-72.3%	0.1177
	Threat	-64.1%	0.3927
	Toxicity	32.5%	0.1177
Commonsense Morality	Ethics	3.4%	0.2427
CrowS-Pairs	Bias	0.0%	0.0677
DecodingTrust	Stereotype	0.0%	0.0677
DynaHate	Hate	10.0%	0.1177
Exaggerated Safety	Safety	2081.6%	0.3854
HASOC (En)	Hate, Offensive	110.2%	0.1177
HateCheck	Hate	6.5%	0.1177
HateEval	Hate	-27.7%	0.1177
HatemojiCheck	Hate	8.0%	0.1177
HateXplain	Hate	255.5%	0.3609
Jigsaw	Identity Hate	1238.1%	0.1311
	Insult	420.0%	0.1177
	Obscene	509.5%	0.1177
	Severe Toxic	672.7%	0.1177
	Threat	*3000.0%*	0.3927
	Toxic	333.9%	0.1177
OpenAI Content Moderation	Harassment	842.9%	0.2427
	Self Harm	837.4%	0.2427
PKU-Alignment-BeaverTails-Eval	Unsafe	277.2%	0.4036
Reddit Content Moderation	Rule Moderation	-20.8%	0.4153
Scruples Anecdotes	Ethics	600.0%	0.2427
Social Bias Inference Corpus (SBIC)	Intentionally Offensive	6.6%	0.1177
	Potentially Offensive	15.0%	0.1177
	Sexually Offensive	-71.7%	0.1177
SWAD	Swear	29.9%	0.1177
ToxiGen	Toxic	20.9%	0.1177
TrustworthyLLM	Safety	46.8%	0.4104
TweetEval	Hate	80.5%	0.1177
	Irony	*3000.0%*	0.2427
	Offensive	19.0%	0.1177
USElectionHate	Hate	1052.9%	0.3337
White Supremacist	Hate	27.0%	0.1311

Figure 1: Heatmap between task semantic similarities and relative performance gains. The semantic similarities are calculated based on Table 9 and Table 10. For visualization, relative gains greater than 3000% are set to 3000% and marked in italic.

Dataset	Task	Phi-3-mini (CT-FT)	Mixtral-8x7B (CT-FT)	**STAND-Guard**
PKU-Alignment BeaverTails	Animal Abuse	0.667	0.713	0.742
	Child Abuse	0.474	0.840	0.815
	Controversial Topics, Politics	0.143	0.412	0.446
	Discrimination, Stereotype	0.674	0.747	0.731
	Drug Abuse, Weapons	0.687	0.744	0.746
	Financial & Property Crime	0.720	0.765	0.744
	Hateful & Offensive Language	0.676	0.671	0.670
	Misinformation	0.000	0.025	0.082
	Non-Violent Unethical Behavior	0.589	0.683	0.655
	Privacy Violation	0.738	0.799	0.800
	Self Harm	0.593	0.710	0.727
	Sexually Explicit	0.547	0.653	0.667
	Terrorism, Organized Crime	0.045	0.125	0.196
	Violence	0.779	0.819	0.800
PKU-Alignment Safe-RLHF	Unsafe	0.828	0.843	0.871
Private	Hate	0.734	0.796	0.827
	Self Harm	0.786	0.814	0.856
	Sexual	0.696	0.808	0.802
	Violence	0.671	0.773	0.745
AVG		0.581	0.671	0.680

Table 13: Impact of model size on F1 scores for fine-tuned models on in-distribution tasks under zero-shot setting. It is an expansion of Table 5.

Dataset	Task	Phi-3-mini (CT-FT)	Mixtral-8x7B (CT-FT)	**STAND-Guard**
Anatomy of Online Hate	Hate	0.306	0.441	0.269
BIG-bench (German)	Gender Inclusive	0.000	0.677	0.708
CallMeSexist	Sexism	0.600	0.741	0.724
Civil-Comments	Insult	0.548	0.597	0.787
	Obscenity	0.124	0.239	0.179
	Severe Toxicity	0.458	0.363	0.530
	Sexually Explicit	0.008	0.073	0.134
	Threat	0.088	0.127	0.163
	Toxicity	0.586	0.628	0.759
Commonsense Morality	Ethics	0.723	0.719	0.735
COVID-HATE	Hate	0.744	0.694	0.695
CrowS-Pairs	Bias	1.000	1.000	1.000
DecodingTrust	Stereotype	1.000	1.000	1.000
DynaHate	Hate	0.614	0.710	0.673
Exaggerated Safety	Safety	0.839	0.886	0.829
GermEval (German) bin	Offensive	0.481	0.612	0.501
GermEval (German) multi	Offensive	0.511	0.677	0.101
HASOC (English)	Hate, Offensive	0.655	0.643	0.679
HASOC (German)	Hate, Offensive	0.416	0.474	0.417
Hate Speech and Offensive Language	Hate, Offensive	0.837	0.719	0.585
Hate Speech towards Foreigners (German)	Hate	0.542	0.610	0.495
HateCheck	Hate	0.856	0.925	0.867
HateEval	Hate	0.267	0.610	0.462
HatemojiCheck	Hate	0.705	0.879	0.836
HateXplain	Hate	0.761	0.801	0.782
Jigsaw (Toxic Comment Classification)	Identity Hate	0.173	0.350	0.281
	Insult	0.395	0.494	0.416
	Obscene	0.422	0.574	0.512
	Severe Toxic	0.070	0.149	0.085
	Threat	0.063	0.328	0.300
	Toxic	0.502	0.582	0.551
Jiminy-Cricket	Ethics	0.607	0.664	0.598
Korean Hate Speech (Korean)	Hate	0.215	0.449	0.240
	Aggressiveness	0.430	0.510	0.382
	Bias	0.733	0.743	0.698
OffComBR3 (Portuguese)	Offensive	0.426	0.555	0.558
OpenAI CM	Harassment	0.384	0.494	0.726
	Hateful	0.657	0.709	0.732
	Self Harm	0.731	0.712	0.928
	Sexual	0.746	0.735	0.475
	Violence	0.857	0.943	0.671
PKU-Alignment-BeaverTails-Eval	Unsafe	0.323	0.417	0.513
Reddit Content Moderation	Rule Moderation	0.134	0.167	0.126
RP-Mod & RP-Crowd (German)	Offensive	0.308	0.249	0.210
Scruples Anecdotes	Ethics	0.419	0.448	0.427
Social Bias Inference Corpus (SBIC)	Intentionally Offensive	0.717	0.721	0.709
	Potentially Offensive	0.715	0.740	0.728
	Sexually Offensive	0.464	0.515	0.195
SWAD	Swear	0.597	0.627	0.539
SWSR (Chinese) bin	Sexism	0.433	0.559	0.504
SWSR (Chinese) multi	Sexism	0.520	0.538	0.528
ToLD-BR (Portuguese)	Offensive	0.560	0.663	0.302
	Homophobia	0.187	0.375	0.140
	Misogyny	0.133	0.214	0.624
	Racism	0.152	0.325	0.215
ToxiGen	Toxic	0.271	0.492	0.601
TrustworthyLLM	Safety	0.603	0.708	0.590
TweetEval	Hate	0.615	0.602	0.556
	Irony	0.766	0.758	0.685
	Offensive	0.472	0.616	0.682
USElectionHate	Hate	0.222	0.487	0.392
White Supremacist	Hate	0.540	0.733	0.739
AVG		0.488	0.577	0.533

Table 14: Impact of model size on F1 scores for fine-tuned models on out-of-distribution tasks under zero-shot setting. It is an expansion of Table 5.