ToVo: Toxicity Taxonomy via Voting

Tinh Son Luong[1*], **Thanh-Thien Le**[2*], **Thang Viet Doan**[3*],
Linh Ngo Van[4†], **Thien Huu Nguyen**[5], **Diep Thi-Ngoc Nguyen**[6†]
[1]Oraichain Labs Inc., US [2]VinAI Research [3]Florida International University
[4]Hanoi University of Science and Technology [5]University of Oregon
[6]VNU University of Engineering and Technology
tinh.ls@orai.io, v.thienlt3@vinai.io, tdoan011@fiu.edu,
linhnv@soict.hust.edu.vn, thien@cs.oregon.edu, ngocdiep@vnu.edu.vn

Abstract

Existing toxic detection models face significant limitations, such as lack of transparency, customization, and reproducibility. These challenges stem from the closed-source nature of their training data and the paucity of explanations for their evaluation mechanism. To address these issues, we propose a dataset creation mechanism that integrates voting and chain-of-thought processes, producing a high-quality open-source dataset for toxic content detection. Our methodology ensures diverse classification metrics for each sample and includes both classification scores and explanatory reasoning for the classifications.

We utilize the dataset created through our proposed mechanism to train our model, which is then compared against existing widely-used detectors. Our approach not only enhances transparency and customizability but also facilitates better fine-tuning for specific use cases. This work contributes a robust framework for developing toxic content detection models, emphasizing openness and adaptability, thus paving the way for more effective and user-specific content moderation solutions.

1 Introduction

Detecting toxicity in text generation to ensure safe interactions between human and Large Language Models (Radford et al., 2019; Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023; Achiam et al., 2023; Team et al., 2023) is a pivotal research challenge. Despite numerous publications, most contributions have focused on releasing toxicity benchmarking datasets (Gehman et al., 2020; Hartvigsen et al., 2022; Luong et al., 2024). Regarding detection mechanisms, the research community still relies heavily on partially or fully closed-source models such as Llama Guard (Inan et al., 2023) and OpenAI Moderations[1] (OAIM). This dependency introduces significant limitations, including a lack of transparency, customization, and reproducibility.

For instance, users cannot fine-tune these models for their own use cases, such as adapting the detection model to address novel forms of toxicity relevant to unique community standards. Additionally, these models often provide no explanation of their evaluation methods, leading to misunderstandings about the model's performance and making it challenging to control quality. Furthermore, the training data for these models is predominantly closed-source, which hampers efforts to customize detection models or reproduce their results, thereby hindering improvement in this crucial area.

To address these shortcomings, we need to make the following observations. First, to address the lack of interpretability, when the detection model processes a sample, the output must include a classification score and an accompanying explanation for the classification. Second, to facilitate user-driven fine-tuning, the original model should be diverse, covering a wide range of toxicity criteria from its inception. To meet these desiderata, we develop a toxic detection dataset creation mechanism via voting and chain of thought.

Overall, our contributions are as follows:

a. We introduce the **To**xicity Taxonomy **Vo**ting (**ToVo**) dataset, a comprehensive resource that categorizes each sample using a diverse selection of metrics from a pool of 42 derived from four different moderation tools. This extensive coverage ensures that the dataset addresses multiple aspects of toxic content detection. Each classification outcome is generated by a set of open-source models and includes an explanatory rationale, providing valuable insights into the reasoning behind each

[*]Equally contributed.
[†]Corresponding authors.

[1]https://platform.openai.com/docs/guides/moderation

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 201–212

Model/Dataset	Metric	Consensus	Metric	Consensus
Llama Guard 2	Child Sexual Exploitation	85.484	Hate	82.353
	Indiscriminate Weapons	92.761	Intellectual Property	94.840
	Non-Violent Crimes	83.279	Privacy	86.688
	Sex-Related Crimes	82.353	Sexual Content	76.440
	Specialized Advice	90.271	Suicide & Self-Harm	97.272
	Violent Crimes	79.630	**Overall**	**86.576**
OAIM	harassment	78,654	harassment/threatening	81.422
	hate	88.090	hate/threatening	85.990
	self-harm	96.677	self-harm/instructions	97.928
	self-harm/intent	96.508	sexual	91.233
	sexual/minors	94.634	violence	89.218
	violence/graphic	90.612	**Overall**	**90.045**
Perspective API	Identity attack	87.158	Insult	75.912
	Profanity	87.100	Severe Toxicity	74.336
	Threat	90.060	Toxicity	72.063
			Overall	**81.013**
BeaverTails	14 other metrics	N/A	**Overall**	**N/A**

Table 1: 42 toxicity metrics available in **ToVo**. The quantity *Consensus* denotes the percentage of agreement between the original toxicity API/model and the voting outcomes from multiple LLMs. ***Note:*** The BeaverTails dataset does not come with a model or API, so the consensus rate cannot be calculated.

classification. This dataset is crucial for developing robust and adaptable toxicity detection models.

b. We leverage the **ToVo** dataset to develop adaptive taxonomy classification models, capable of operating effectively with both predefined and user-tailored metrics. To demonstrate their efficiency, we benchmark our models against leading moderation tools such as PerspectiveAPI[2], OAIM, and Llama Guard 2 on their respective predefined metrics. Additionally, we conduct rigorous Out-of-Domain benchmarking using an evaluation dataset with metrics unrelated to toxicity, showcasing the versatility and robustness of our models.

2 Dataset Constructions

2.1 Dataset Sourcing

To create the **ToVo** dataset, we first compile a collection of prompts paired with their respective responses. We begin by extracting prompts from the `chat-lmsys-1M` (Zheng et al., 2023) dataset, consisting of 1 million conversations between multiple LLMs and their users. Given that this is a general-purpose dataset, many of its sentences do not contain any toxicity. Therefore, we use HateBERT (Caselli et al., 2020) to perform a preliminary filtering process, retaining only prompts whose responses exceed a predefined threshold of toxicity probability; this practice has been previously done by Luong et al. (2024). Subsequently, we randomly select 10,000 prompts from this filtered subset, including 5,000 prompts from single-turn conversations and 5,000 from multi-turn conversations. From these, we obtain responses by prompting open-source models such as Mistral-Instruct (Jiang et al., 2023) and Zephyr (Tunstall et al., 2023).

2.2 Classification Label

To establish a gold-labeled taxonomy dataset, we implement a rigorous voting procedure. One of our motivations is to make our dataset similar to the data users might use to fine-tune the model, which can be heterogeneous in terms of the number and type of toxicity metrics in each sample. As a result, we collect a pool of 42 predefined toxicity metrics from Llama Guard 2 - MLCommons (Vidgen et al., 2024), OAIM, Perspective API, and BeaverTails (Ji et al., 2024). For each sample in the filtered subset, we randomly select 1 to 6 metrics to classify the sample on. Subsequently, three out of six open-source LLMs, which are listed in Appendix A.1 are randomly selected to vote on whether the sample is positive for each of its selected metrics. Criteria for selection include the model's ability to produce sufficiently accurate classifications while avoiding excessively stringent criteria that might

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

prematurely block prompts.

Following this selection, classification results are generated for each chosen model based on the previously selected metrics. To enhance the interpretability of these results, a Chain-of-Thought (Wei et al., 2022) prompting technique is applied during the generation process. This method facilitates a more nuanced and comprehensive understanding of the classification outcomes.

2.3 Classification Rationale

As mentioned earlier, three LLMs determine the classification labels for each sample. To select whose rationale would be used as the primary explanation, we engaged in a ranking procedure for each of the six predetermined open-source models, assessing their consensus rates relative to other models. The consensus rate quantifies the level of agreement between the classification outputs of the focal model and the aggregate classifications generated by all selected models.

After gathering these consensus rates, we used the rationale from the model with the highest consensus rate among those that agreed with the majority classification for each sample. This approach ensured that we chose the most consistent and harmonized classification outcome across models, helping to mitigate discrepancies.

3 Experiments

3.1 Dataset Alignment Evaluation

We evaluate the alignment of our dataset with other moderation APIs and models, including Llama Guard 2, OpenAI moderation, and the Perspective API. Specifically, for each metric, we measure the *consensus rate*, which is the percentage of agreement between the gold labels obtained via our voting process and the outputs from the original API/model. The results are presented in Table 1. OAIM shows the highest overall consensus, indicating it is the most aligned with our dataset, while Perspective API has the lowest, suggesting that it might benefit from further alignment with our voting-based gold labels.

Overall, the observed high consensus rates demonstrate a high level of agreement between our gold labels and the outputs from Llama Guard 2 and OAIM, whereas Perspective API shows more variability. This suggests that our voting process produces reasonable and consistent gold labels for toxicity classifications, particularly among metrics with high consensus rates, as predictions should align closely with the reference models rather than being arbitrary. Additionally, this method is scalable, enabling developers to create their own datasets tailored to specific content and metrics.

3.2 Baseline Model Training & Evaluation

3.2.1 Training configuration

We trained two baseline models using the pretrained `Mistral-Hermes-2-Pro` [3] from NousResearch with 10,000 samples derived from the voting process, utilizing `transformers` library (Wolf et al., 2019) and LoRA (Hu et al., 2022). One model outputs reasoning for each classification, while the other provides only the classification results. Both model variations—reasoning and non-reasoning—were fine-tuned using LoRA, with a rank of 16 and an α of 16. The training was conducted on a single A100 GPU (40GB memory) with a batch size of 4 and 8 gradient accumulation steps, resulting in a global batch size of 32. The learning rate was set to $1e-4$, and the models were trained over 2 epochs. The prompt templates used for training the reasoning and non-reasoning models are detailed in Appendix A.3

3.2.2 Toxicity Taxonomy Evaluation

To evaluate our baseline models, we test them on a set of 2,322 samples with toxicity-related metrics. We compare the alignment of our baseline models with other moderation tools. The results are presented in Table 2 and Appendix A.4. It is noteworthy that lower consensus rates in certain metrics when comparing our models with other tools (e.g., "Sexual Content" for Llama Guard 2 and "Toxicity" for Perspective API) do not necessarily suggest areas where our classification criteria might need refinement. Instead, they could indicate that while the metrics are the same, the level of toxicity tolerance that determines whether a sample should be labeled as positive differs between our model and the original API. We also benchmark our models with the labels obtained from the voting process; the details are presented in Appendix A.4.

3.2.3 Out-of-Domain Taxonomy Evaluation

Additionally, we construct an Out-of-Domain (OOD) test set using metrics unrelated to toxicity. This experiment aims to evaluate the generalization

[3] https://huggingface.co/NousResearch/Hermes-2-Pro-Mistral-7B

Model/Dataset	Consensus	Model/Dataset	Consensus	Model/Dataset	Consensus
Llama Guard 2	91.187 *92.264*	OAIM	89.981 *90.625*	Perspective AI	77.557 *80.871*

Table 2: The consensus rate between the outputs from our models and the gold labels obtained via the original API/model. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	***Consensus-R***
Out-of-Domain	Educational Content	90.751	*92.741*
	Health and Wellness	98.057	*99.297*
	Science and Technology	37.117	*76.223*
	Arts and Culture	96.516	*97.527*
	Travel and Adventure	97.245	*98.087*
	Personal Development	98.590	*99.167*
	Cooking and Recipes	91.887	*95.009*
	Gardening and Horticulture	80.304	*84.589*
	Fitness and Exercise	64.416	*86.539*
	Financial Literacy	94.662	*97.312*
	Overall	**84.689**	***92.536***

Table 3: The consensus rate between the outputs from our models and the gold labels obtained via our voting process on the Out-of-Domain metrics. *Italicized* values denote results from the model trained with reasoning.

of the baseline models and their zero-shot performance on user-custom metrics.

In the OOD test set, there are 10 metrics that are unrelated to the toxicity metrics used in the training set. Using the specific definitions of these metrics, which are detailed in Appendix A.2, we apply the construction process described in Section 2 to create a test set comprising 1741 samples. To evaluate our models, we compare the consensus rate between our models and the gold labels according to the voting process described in Section 2. The result are available in Table 3.

Both models achieve high consensus rates for several OOD metrics, especially the reasoning model: it outperforms its non-reasoning counterpart across all metrics. Additionally, some metrics exhibit more variability in consensus rates between the standard and reasoning models. The "Fitness and Exercise" metric, for example, shows a notable improvement from 64.416% to 86.539% with the reasoning model, suggesting that certain categories benefit significantly from the additional interpretability provided by reasoning. The overall consensus rate across all OOD metrics is 84.689% for the non-reasoning model and 92.536% for the reasoning model. This overall improvement underscores the value of the reasoning approach in achieving more reliable and consistent classification outcomes for high-novelty metrics.

This experiment demonstrates that our models, particularly those incorporating reasoning, are highly adaptable and effective in classifying content across a wide range of domains. This adaptability is crucial for real-world applications, enhancing the reliability of the models in handling diverse and dynamic content types.

4 Conclusions

In this paper, to address significant limitations in existing toxic content detection models, we introduce the **To**xicity Taxonomy **Vo**ting (**ToVo**) dataset, developed through a rigorous voting mechanism and Chain-of-Thought prompting to ensure high-quality, explainable classification outcomes.

Utilizing **ToVo**, we train two taxonomy models that perform exceptionally well on toxicity-related metrics in the evaluation dataset. These models not only align closely with the gold labels generated by the voting process but also demonstrate a high level of consensus with other moderation tools. Furthermore, our reasoning model achieves exceptional results on the Out-of-Domain test set, affirming its adaptability to user-specific custom metrics. This underscores the model's potential for fine-tuning in diverse application scenarios.

Our work introduces a novel method for creating customized moderation tools effortlessly, using an automated process that combines Voting and

Chain-of-Thought techniques. By fostering safer human-LLM interactions and empowering users with customizable moderation tools, we hope this work can pave the way for creating a safer and more inclusive digital environment.

Limitations & Future Works

While the voting process with Chain-of-Thought prompting is efficient, generating large amounts of data remains time-consuming and labor-intensive. Additionally, utilizing large language models (LLMs) for taxonomy purposes results in slow inference speeds. This slowdown is attributed to the substantial size of the models (7 billion parameters) and the inclusion of reasoning in some responses, which further hampers processing speed.

Moreover, the current version of our taxonomy only supports binary classification. This binary approach can sometimes be insufficient, as classifying metrics solely as 0 or 1 may not accurately capture the nuances of their potential unsafety.

In the future, we aim to extend our research beyond large language models (LLMs) to widen its range of applications. Specifically, we plan to apply our toxicity detection framework to human-human interactions, web content, and online forums. By doing so, we must enhance the robustness and versatility of our models, ensuring they can effectively handle diverse and dynamic contexts of toxic content across various platforms. This expansion will also involve refining our taxonomy to support more nuanced classifications, enabling more accurate and context-sensitive moderation. Ultimately, our goal is to contribute to safer and more inclusive digital environments through adaptable toxicity detection solutions.

Acknowledgements

This research was partially supported by NSF Grant #2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

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*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *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*.

Tinh Son Luong, Thanh-Thien Le, Linh Ngo Van, and Thien Huu Nguyen. 2024. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

"Teknium", "theemozilla", "karan4d", and "huemin_art". Nous hermes 2 mixtral 8x7b dpo.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. 2024. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

A Appendix

A.1 Voting Models

The six LLMs we use during our Voting Process described in Section 2.2 are:

- `deepseek-llm-67b-chat` (DeepSeek-AI, 2024),
- `dolphin-2.5-mixtral-8x7bv`[4],
- `Nous-Hermes-2-Mixtral-8x7B-DPO` ("Teknium" et al.),
- `WizardLM-2-8x22B`[5],
- `Yi-34B-Chat` (AI et al., 2024), and
- `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024).

A.2 Out-of-Domain Test Set

In the Out-of-Domain Test Set, there are 10 metrics that are unrelated to the toxicity metrics used in the training set. The specific descriptions of these 10 metrics are presented in Table 4. Using these definitions, we apply the construction process described in Section 2 to create a test set comprising 1741 samples.

A.3 Prompt Templates

The prompt templates used for training the reasoning and non-reasoning models are detailed in Figures 1 and 2, respectively.

[4]https://huggingface.co/cognitivecomputations/dolphin-2.5-mixtral-8x7b

[5]https://huggingface.co/alpindale/WizardLM-2-8x22B

A.4 Additional Experimental Results

Table 5, 6, and 7 present the performance of our models, using the gold labels from the original toxicity detection API/model.

Table 8, 9, 10, and 11 present the performance of our models, using the gold labels from our voting process.

A.5 Experiments on BeaverTails-30K test set

To further evaluate our models against others such as MD Judge and Llama Guard 2, we conduct experiments using the BeaverTails-30K test set. It is important to note that while our model is trained on a dataset incorporating metrics similar to those used in BeaverTails, it is not trained directly on the BeaverTails dataset.

The results of our experiments are presented in Table 12. From here we can make the following observations:

- Solely relying on the metric's description to classify content as safe or unsafe is insufficient. It is crucial to also consider the degree to which the content may be deemed unsafe.
- For instance, Some prompts in BeaverTails involve "how to prank..." scenarios. While most of these are classified as Non-Violence Unethical Behaviors, which is safe, both our models and the voting results from other models suggest that such content has the potential to be unsafe. However, without specific details about the prank, it is challenging to classify the content as unsafe within the BeaverTails framework. This suggests that the metric descriptions in BeaverTails may not fully capture the nuances of their manual labeling process.
- We attempted to broaden the scope of the metric descriptions to better classify potential cases as unsafe. This adjustment resulted in a slight improvement in our model's performance on the BeaverTails-30K test set. Consequently, we believe that obtaining metric descriptions that more closely align with BeaverTails' labeling criteria could significantly enhance the accuracy of our models or others.

Metric	Description
Educational Content	This includes content designed to educate and inform users on a variety of topics, such as academic subjects, professional skills, hobbies, and general knowledge. It aims to provide learning opportunities and improve understanding.
Health and Wellness	This refers to content that promotes physical, mental, and emotional well-being. It includes fitness tips, nutritional advice, mental health resources, and general wellness information.
Science and Technology	This category encompasses content related to scientific discoveries, technological advancements, and innovations. It includes research findings, tech reviews, and discussions on scientific topics.
Arts and Culture	This includes content that explores various forms of art and cultural expressions. It covers topics such as visual arts, music, literature, theater, and cultural traditions from around the world.
Travel and Adventure	This pertains to content that inspires and informs about travel destinations, experiences, and adventures. It includes travel guides, adventure stories, and tips for travelers.
Personal Development	This category includes content aimed at personal growth and self-improvement. It covers topics such as goal setting, productivity, motivation, and skills development.
Cooking and Recipes	This includes content related to culinary arts, recipes, cooking tips, and food preparation techniques. It aims to inspire and guide individuals in creating delicious meals.
Gardening and Horticulture	This pertains to content about gardening practices, plant care, landscaping, and horticultural techniques. It includes advice on growing flowers, vegetables, and maintaining gardens.
Fitness and Exercise	This includes content focused on physical fitness, workout routines, exercise techniques, and sports. It aims to promote an active lifestyle and provide guidance on maintaining physical health.
Financial Literacy	This category includes content that educates individuals on managing finances, budgeting, investing, and achieving financial goals. It aims to improve financial understanding and decision-making.

Table 4: Descriptions of the Out-of-Domain metrics.

You are a content classifier tasked with analyzing input content and determining if it falls within pre-defined metrics.
Don't forget to add reasoning in order to explain your classification result. Also do not add metrics which not included in the pre-defined metrics.
The input content:
"**<CONTENT>**"
Pre-defined Metrics:
<METRICS>
Output:

Figure 1: Prompt template for the training of the reasoning model.

You are a content classifier tasked with analyzing input content and determining if it falls within pre-defined metrics.
You must not explain your classification result. Also do not add metrics which not included in the pre-defined metrics.
The input content:
"**<CONTENT>**"
Pre-defined Metrics:
<METRICS>
Output:

Figure 2: Prompt template for the training of the non-reasoning model.

Model/Dataset	Metric	Consensus	*Consensus-R*
Llama Guard 2	Child Sexual Exploitation	90.659	*90.643*
	Indiscriminate Weapons	97.959	*97.850*
	Non-Violent Crimes	89.474	*90.798*
	Sex-Related Crimes	82.486	*87.500*
	Specialized Advice	96.809	*97.802*
	Violent Crimes	89.286	*91.926*
	Hate	86.700	*87.113*
	Intellectual Property	99.471	*96.757*
	Privacy	93.252	*96.795*
	Sexual Content	77.515	*77.640*
	Suicide & Self-Harm	97.382	*98.387*
	Overall	91.187	*92.264*

Table 5: The consensus rate between the outputs from our models and the gold labels obtained via the original API/model on the metrics from Llama Guard 2. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	*Consensus-R*
OAIM	Harassment	73.714	*78.049*
	Hate	88.095	*87.180*
	Self-Harm	98.305	*97.647*
	Self-Harm/Intent	100.000	*99.492*
	Sexual/Minors	93.367	*95.676*
	Violence/Graphic	89.560	*91.954*
	Harassment/Threatening	86.473	*85.714*
	Hate/Threatening	84.536	*84.946*
	Self-Harm/Instructions	95.238	*94.479*
	Sexual	87.817	*90.426*
	Violence	91.848	*90.173*
	Overall	89.981	*90.625*

Table 6: The consensus rate between the outputs from our models and the gold labels obtained via the original API/model on the metrics from OpenAI Moderation. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	*Consensus-R*
Perspective AI	Identity attack	91.667	*93.902*
	Profanity	84.211	*86.264*
	Threat	89.063	*90.000*
	Insult	68.023	*73.913*
	Severe Toxicity	67.172	*72.251*
	Toxicity	65.946	*69.663*
	Overall	77.557	*80.871*

Table 7: The consensus rate between the outputs from our models and the gold labels obtained via the original API/model on the metrics from Perspective API. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	*Consensus-R*
Llama Guard 2	Child Sexual Exploitation	96.133	*97.647*
	Indiscriminate Weapons	100.000	*99.462*
	Non-Violent Crimes	89.412	*90.124*
	Sex-Related Crimes	93.220	*90.533*
	Specialized Advice	97.340	*96.721*
	Violent Crimes	94.048	*95.652*
	Hate	94.555	*92.228*
	Intellectual Property	99.471	*99.460*
	Privacy	93.827	*93.548*
	Sexual Content	97.024	*98.125*
	Suicide & Self-Harm	97.906	*97.861*
	Overall	95.833	*95.657*

Table 8: The consensus rate between the outputs from our models and the gold labels obtained via our voting process on the metrics from Llama Guard 2. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	*Consensus-R*
OAIM	Harassment	90.857	*89.091*
	Hate	88.691	*91.667*
	Self-Harm	98.864	*99.412*
	Self-Harm/Intent	99.034	*98.469*
	Sexual/Minors	96.939	*96.216*
	Violence/Graphic	93.407	*97.126*
	Harassment/Threatening	87.864	*88.205*
	Hate/Threatening	86.598	*88.172*
	Self-Harm/Instructions	99.405	*99.387*
	Sexual	97.449	*96.791*
	Violence	93.478	*93.103*
	Overall	93.860	*94.311*

Table 9: The consensus rate between the outputs from our models and the gold labels obtained via our voting process on the metrics from Open AI Moderation. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	*Consensus-R*
Perspective AI	Identity attack	93.413	*93.865*
	Profanity	92.105	*94.506*
	Threat	92.188	*92.265*
	Insult	83.140	*83.333*
	Severe Toxicity	87.374	*87.958*
	Toxicity	85.946	*84.358*
	Overall	89.040	*89.414*

Table 10: The consensus rate between the outputs from our models and the gold labels obtained via our voting process on the metrics from Perspective API. *Italicized* values denote results from the model trained with reasoning.

Model/Dataset	Metric	Consensus	*Consensus-R*
BeaverTails	Animal Abuse	99.490	*99.465*
	Child Abuse	97.727	*96.450*
	Controversial Topics, Politics	85.567	*86.631*
	Discrimination, Stereotype, Injustice	89.618	*88.571*
	Drug Abuse, Weapons, Banned Substance	99.320	*99.254*
	Financial Crime, Property Crime, Theft	98.750	*100.000*
	Hate Speech, Offensive Language	89.130	*85.876*
	Misinformation Regarding Ethics, Laws, and Safety	93.258	*95.210*
	Non-Violent Unethical Behavior	85.807	*81.879*
	Privacy Violation	95.313	*94.022*
	Self-Harm	98.438	*98.387*
	Sexually Explicit, Adult Content	96.354	*97.778*
	Terrorism, Organized Crime	99.435	*99.408*
	Violence, Aiding and Abetting, Incitement	91.038	*91.584*

Table 11: The consensus rate between the outputs from our models and the gold labels obtained via our voting process on the metrics from BeaverTails. *Italicized* values denote results from the model trained with reasoning.

Model	Metrics	Unsafe F1
Reasoning	Default BeaverTails Metrics	0.4104
Reasoning	Custom BeaverTails Metrics	0.4164
Voting (200 samples)	Default BeaverTails Metrics	0.4131

Table 12: Performance of Voting process and our reasoning model on BeaverTails-30K test set