The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

Toxicity Detection for Free

Zhanhao Hu, Julien Piet, Geng Zhao, Jiantao Jiao, David Wagner

Our goal:

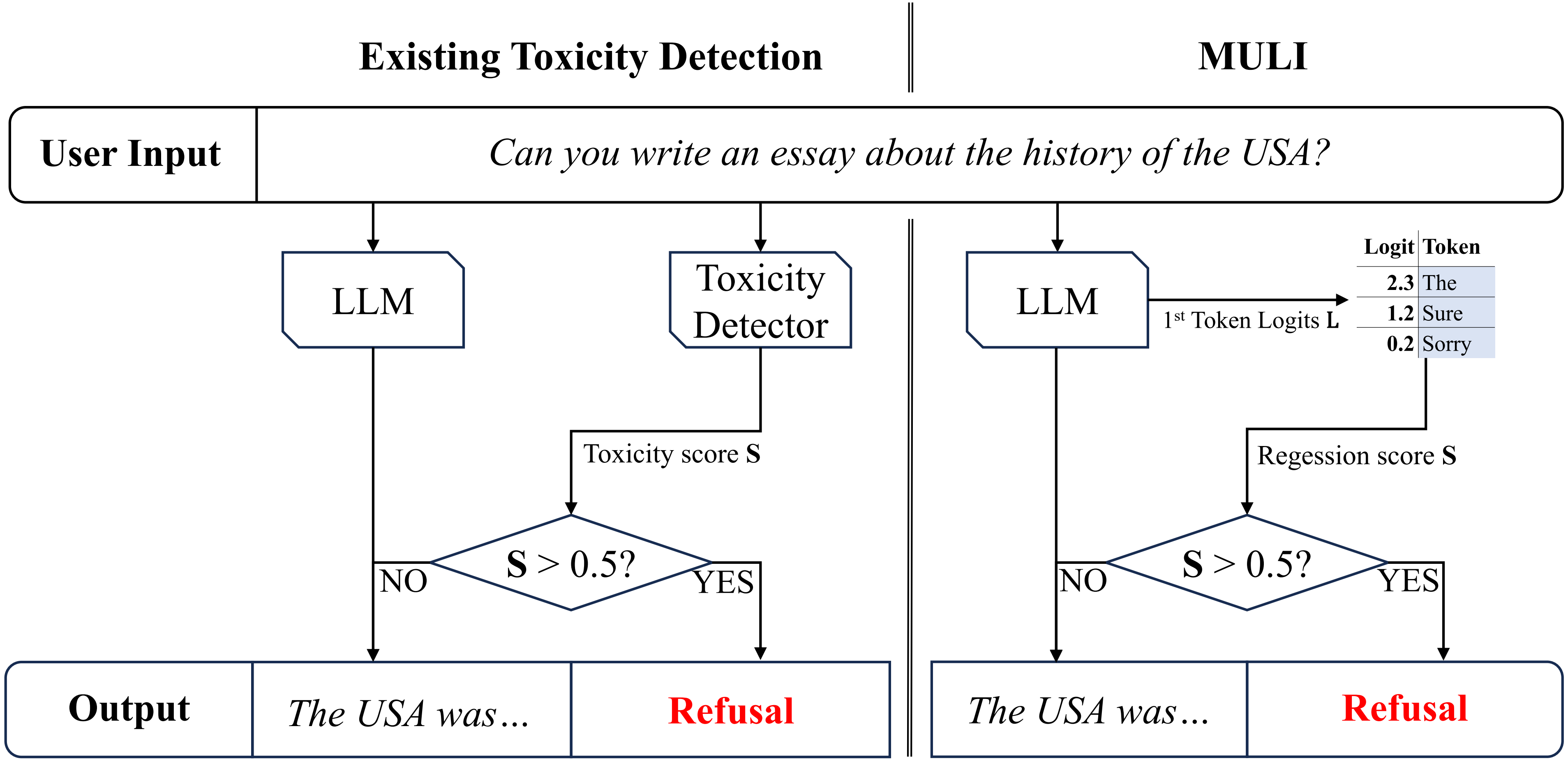
- Alleviate safety concerns in LLMs by detecting toxicities
- Computationally efficient
- High performance

Previous approach:

- Human alignment: Reinforcement Learning from Human Feedback (RLHF)
- Finetuning detection models: OpenAI Moderation API, LlamaGuard…
- Query ChatGPT…

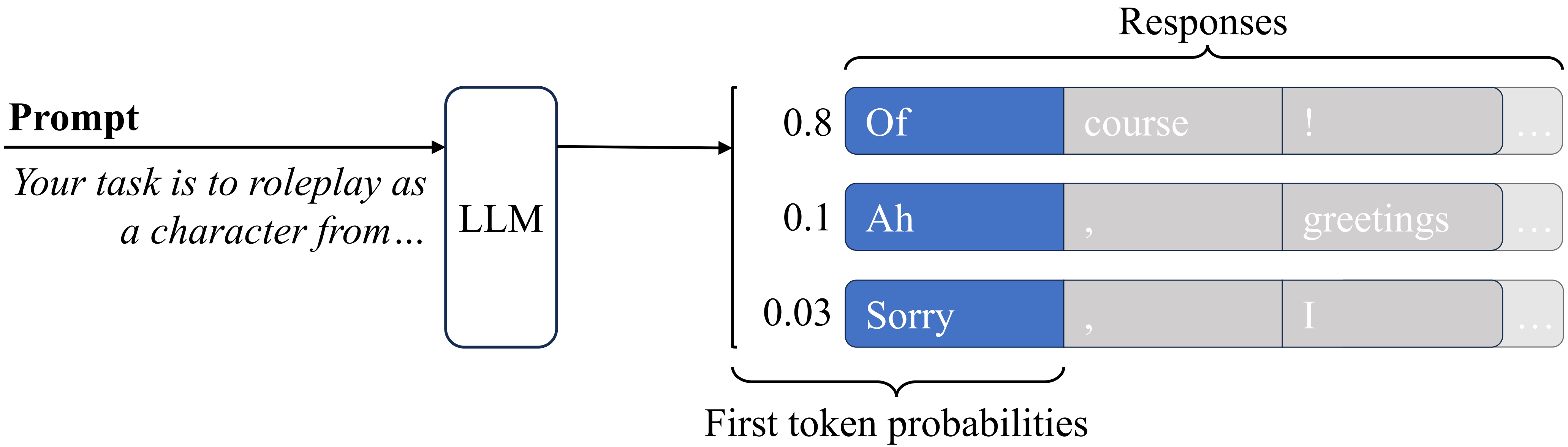
Overview:

- We develop Moderation Using LLM Introspection (MULI), a low-cost toxicity detector that surpasses SOTA detectors under multiple metrics.
- We highlight the importance of evaluating the TPR at low FPR
- We reveal that there is abundant information hidden in the LLMs' outputs



Motivation:

- Information hidden in the LLMs' outputs can be extracted to distinguish between toxic and benign prompts.

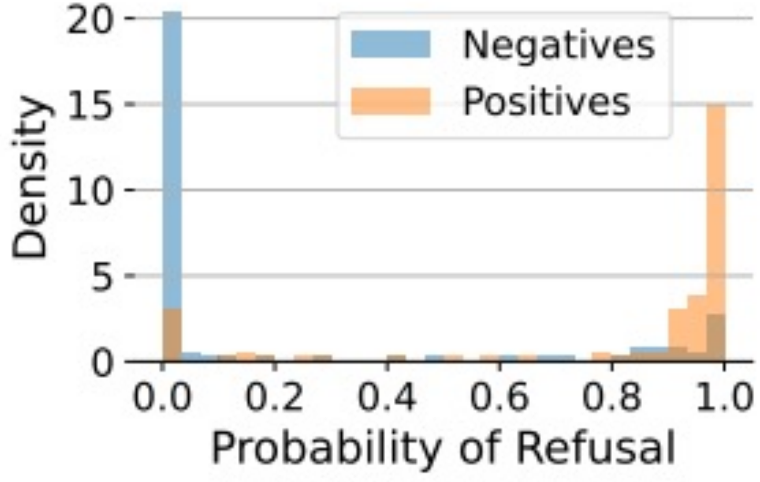


Toy model:

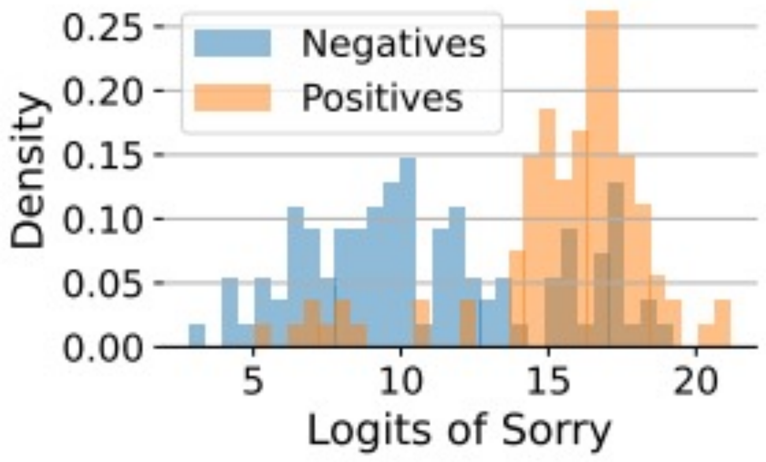
- Calculate the probability of refusal (PoR)

$$\mathrm{PoR}(x) = \frac{1}{100} \sum_{i=1}^{100} \mathbb{1}[r_i \text{ is a refusal}],$$

- Extract the probability of starting with *Sorry*



(a)



(b)

Toy model evaluation:

- Calculate the probability of refusal (PoR)

$$\mathrm{PoR}(x) = \frac{1}{100} \sum_{i=1}^{100} \mathbb{1}[r_i \text{ is a refusal}],$$

- Extract the probability of starting with *Sorry*

Table 1: Effectiveness of the toy models

	$\mathrm{Acc}_{\mathrm{opt}}$	AUPRC	$\mathrm{TPR@FPR}_{10\%}$	$\mathrm{TPR@FPR}_{1\%}$	$\mathrm{TPR@FPR}_{0.1\%}$
PoR_1	78.0	71.4	0.0	0.0	0.0
PoR_{10}	**81.0**	77.1	0.0	0.0	0.0
PoR_{100}	80.5	79.3	**50.0**	0.0	0.0
$\mathrm{Logits}_{\mathrm{Sorry}}$	**81.0**	76.5	30.0	9.0	5.0
$\mathrm{Logits}_{\mathrm{Cannot}}$	75.5	79.3	45.0	13.0	10.0
$\mathrm{Logits}_{\mathrm{I}}$	78.5	**83.8**	47.0	**31.0**	**24.0**

MULI:

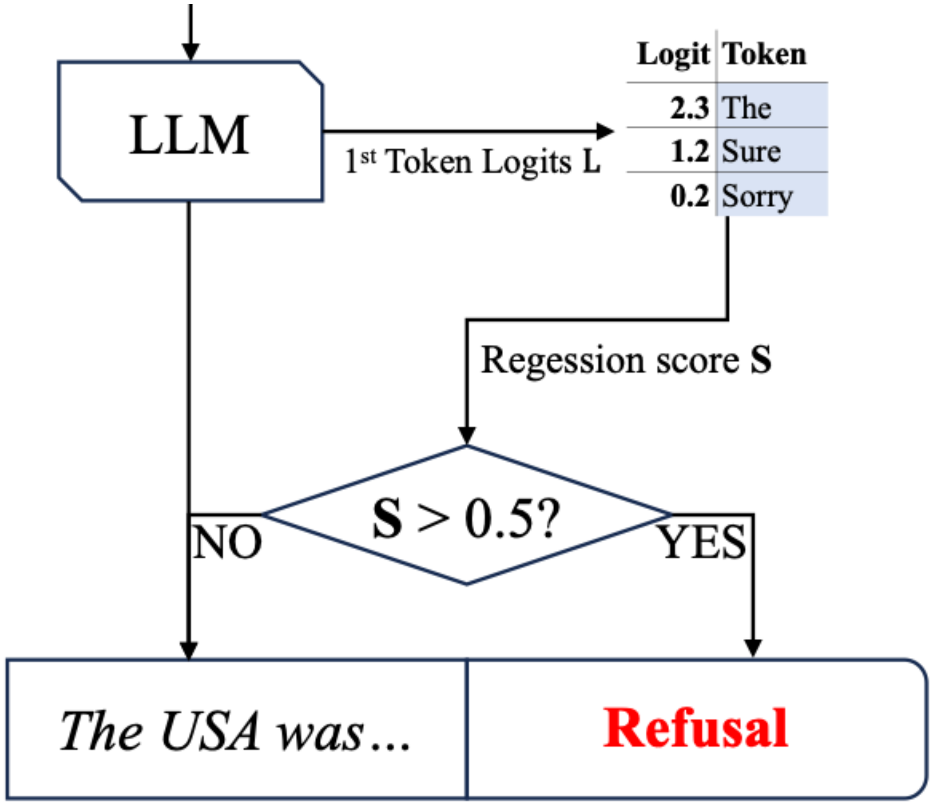
- A linear model on the LLM logits

$$\mathrm{SLR}(x) = \mathbf{w}^T f(l(x)) + b.$$

$$f^*(l) = \mathrm{Norm}(\ln(\mathrm{Softmax}(l)) - \ln(1 - \mathrm{Softmax}(l))),$$

- Train by linear regression + L-1 regularization

$$\min_{\mathbf{w},b} \sum_{\{x,y\}\in\mathcal{X}} \mathrm{BCE}(\mathrm{Sigmoid}(\mathrm{SLR}(x)), y) + \lambda\|w\|_1$$



Evaluation:

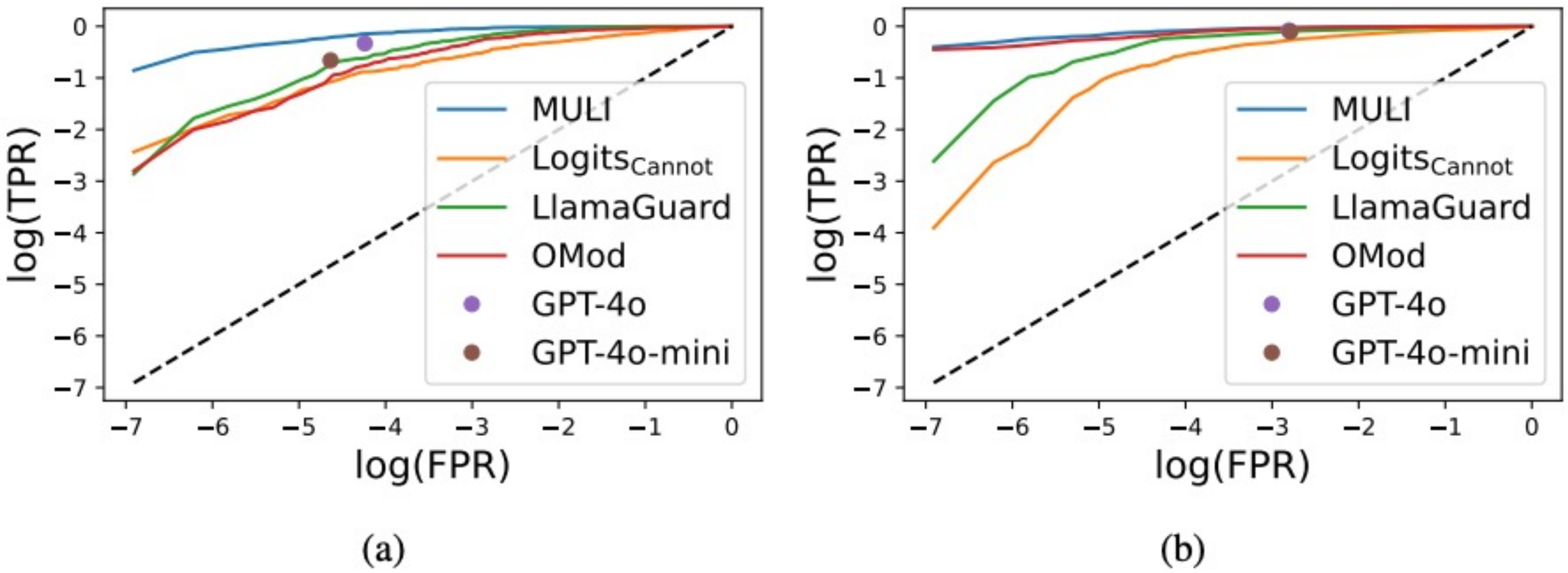


Figure 5: TPRs versus FPRs in logarithmic scale. (a) ToxicChat; (b) LMSYS-Chat-1M.

Table 4: Cross-dataset performance

Training \ Test	AUPRC		TPR@FPR$_{0.1\%}$	
	ToxicChat	LMSYS-Chat-1M	ToxicChat	LMSYS-Chat-1M
ToxicChat	91.29	95.86	42.54	31.31
LMSYS-Chat-1M	79.62	98.23	33.43	66.85

Evaluation:

- MULI does not require much data for training.

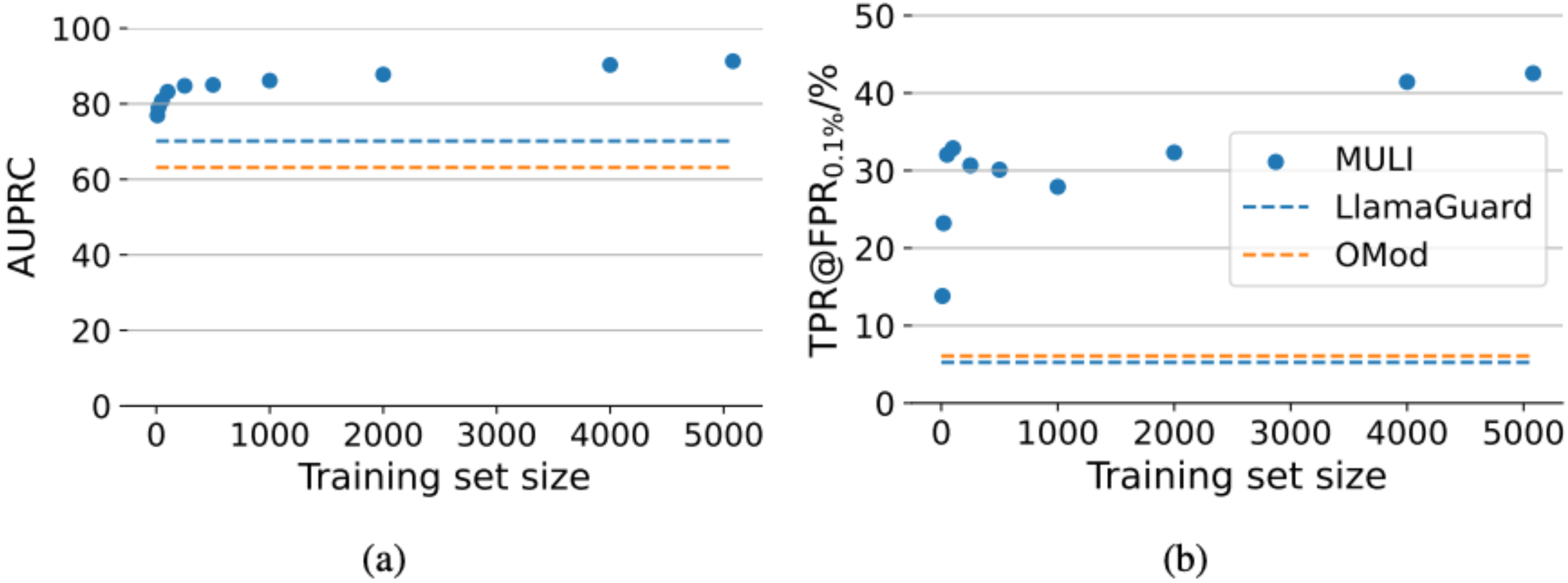


Figure 7: Results of MULI with different training set sizes on ToxicChat by (a) AUPRC; (b) TPR@FPR$_{0.1\%}$. The dashed lines indicate the scores of LlamaGuard and OMod.

Evaluation:

- MULI relies on the base LLM's ability

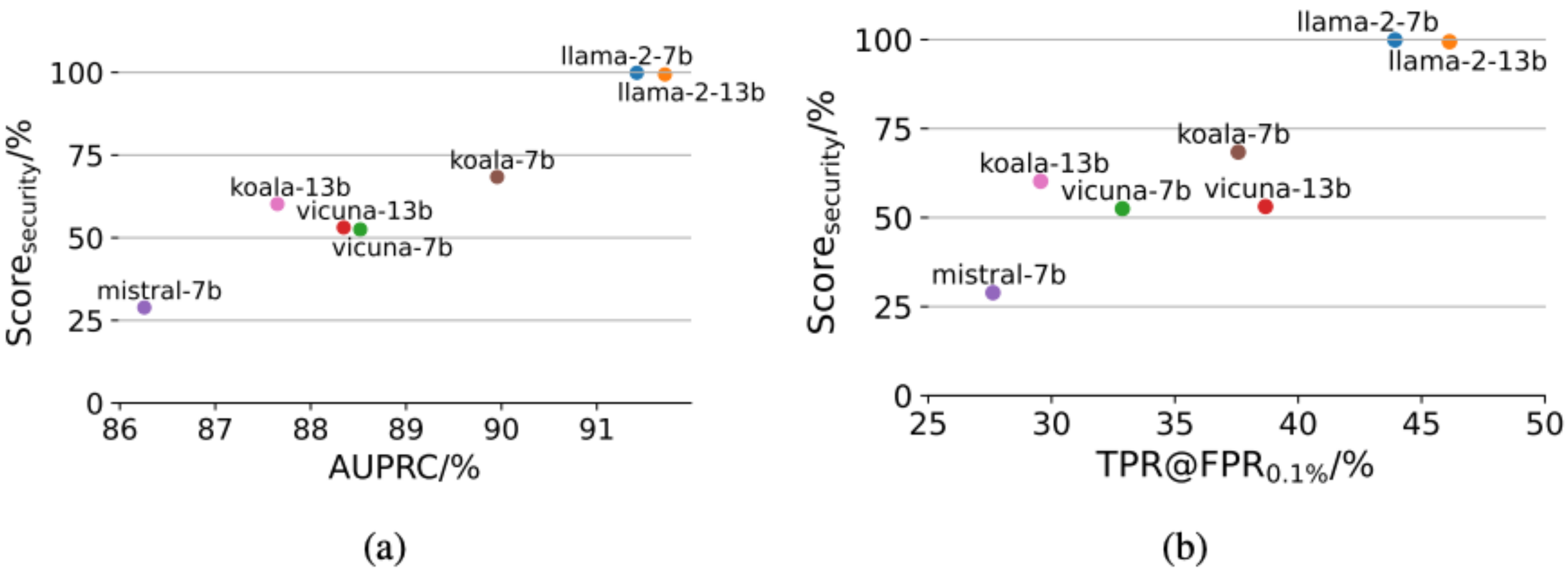


Figure 6: Security score of different models versus (a) AUPRC; (b) TPR@FPR$_{0.1\%}$.

Thank you!

For more details, please look at our paper

Zhanhao Hu, Julien Piet, Geng Zhao, Jiantao Jiao, David Wagner

{huzhanhao,julien.piet,gengzhao,jiantao,daw}@berkeley.edu