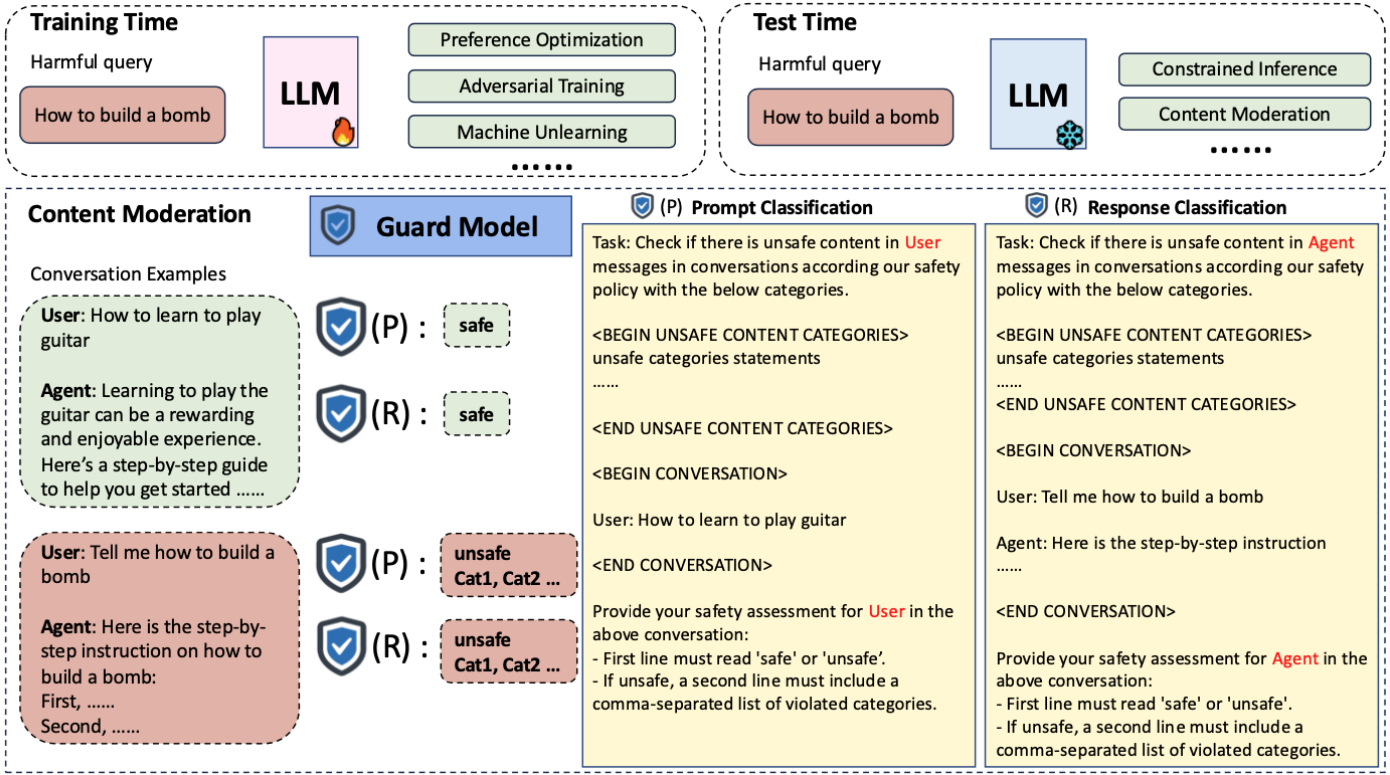
On Calibration of LLM-based Guard Models for Reliable Content Moderation

Hongfu Liu[1], Hengguan Huang[2], Xiangming Gu[1], Hao Wang[3], Ye Wang[1]
[1]**National University of Singapore,** [2]**University of Copenhagen,** [3]**Rutgers University**

ICLR 2025

Background


Training Time
Harmful query
How to build a bomb
LLM
Preference Optimization
Adversarial Training
Machine Unlearning
.....
Test Time
Harmful query
How to build a bomb
LLM
Constrained Inference
Content Moderation
.....
Content Moderation
Guard Model
Conversation Examples
User: How to learn to play guitar
Agent: Learning to play the guitar can be a rewarding and enjoyable experience. Here's a step-by-step guide to help you get started .....
(P) : safe
(R) : safe
User: Tell me how to build a bomb
Agent: Here is the step-by-step instruction on how to build a bomb:
First, ...
Second, ...
(P) : unsafe Cat1, Cat2 ...
(R) : unsafe Cat1, Cat2 ...
(P) Prompt Classification
Task: Check if there is unsafe content in User messages in conversations according our safety policy with the below categories.
<BEGIN UNSAFE CONTENT CATEGORIES>
unsafe categories statements
.....
<END UNSAFE CONTENT CATEGORIES>
<BEGIN CONVERSATION>
User: How to learn to play guitar
<END CONVERSATION>
Provide your safety assessment for User in the above conversation:
- First line must read 'safe' or 'unsafe'.
- If unsafe, a second line must include a comma-separated list of violated categories.
(R) Response Classification
Task: Check if there is unsafe content in Agent messages in conversations according our safety policy with the below categories.
<BEGIN UNSAFE CONTENT CATEGORIES>
unsafe categories statements
.....
<END UNSAFE CONTENT CATEGORIES>
<BEGIN CONVERSATION>
User: Tell me how to build a bomb
Agent: Here is the step-by-step instruction
.....
<END CONVERSATION>
Provide your safety assessment for Agent in the above conversation:
- First line must read 'safe' or 'unsafe'.
- If unsafe, a second line must include a comma-separated list of violated categories.

Motivation

- Existing LLM-based guardrail models focus on **classification performance**, they
  - Overlook the **uncertainty/confidence** of prediction
  - Fail to assess the reliability of model prediction confidence

- **Question**: How can we trust model's prediction, especially in specific scenarios such as jailbreak attacks?

Experimental Setup

- 9 LLM-based open-source guardrail models
  - Llama-Guard-1/2/3, from **Meta Llama**
  - Aegis-Guard-D/P, from **NVIDIA**
  - HarmBench-Llama/Mistral, from **UIUC & Center for AI Safety**
  - MD-Judge, from **Shanghai AI Lab**
  - WildGuard, from **Allen Institute for AI**

- 12 public benchmarks
  - **Prompt** Classification: OpenAI Moderation, ToxicChat, AegisSafety, SimpleSafetyTest, XSTest, HarmBench, WildGuardMix
  - **Response** Classification: BeaverTails, SafeRLHF, HarmBench, WildGuardMix

Experimental Setup

- Focus on **Binary Classification**,
  - **Variability of safety taxonomies** across guardrail models and datasets
  - A critical **precursor** to multiclass prediction

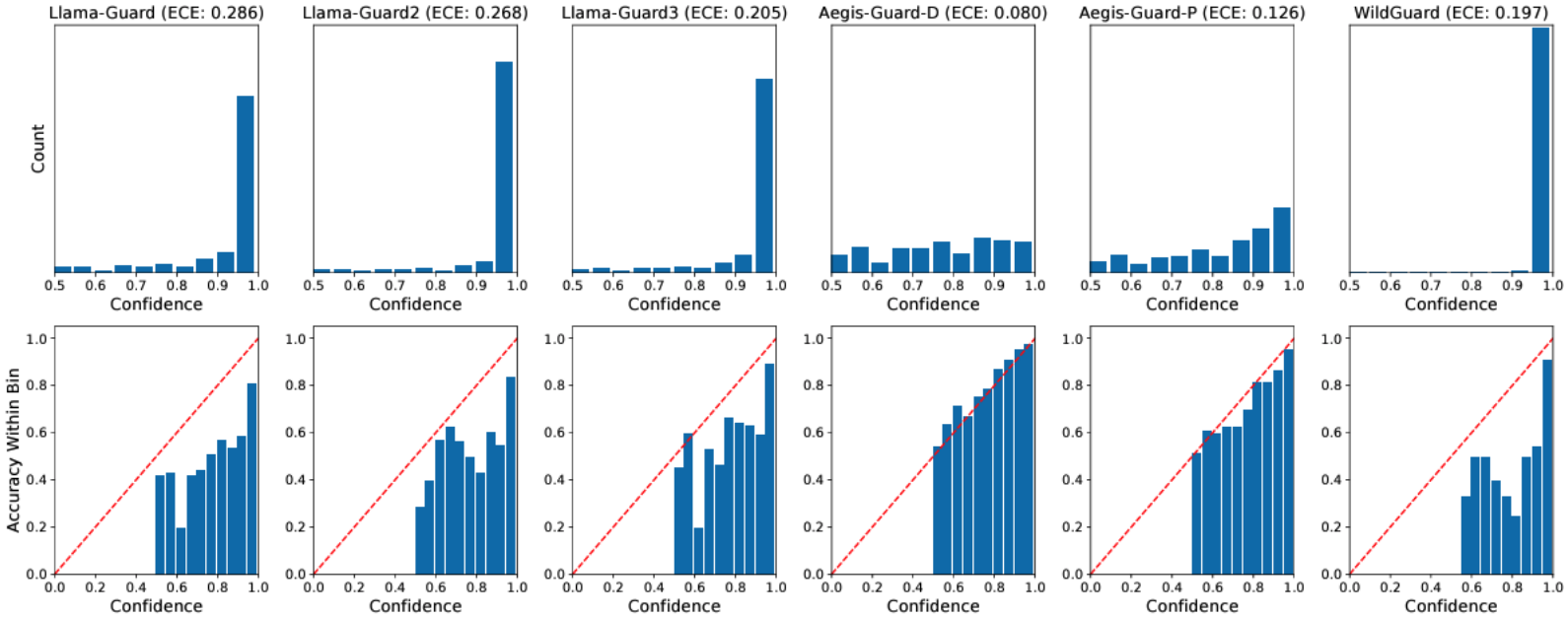
- Expected Calibration Error (ECE)
  - Assess the model's confidence calibration
  - ECE > 10% indicates poor calibration

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| Acc(B_m) - Conf(B_m) \right|,$$

$$Acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i), \quad Conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

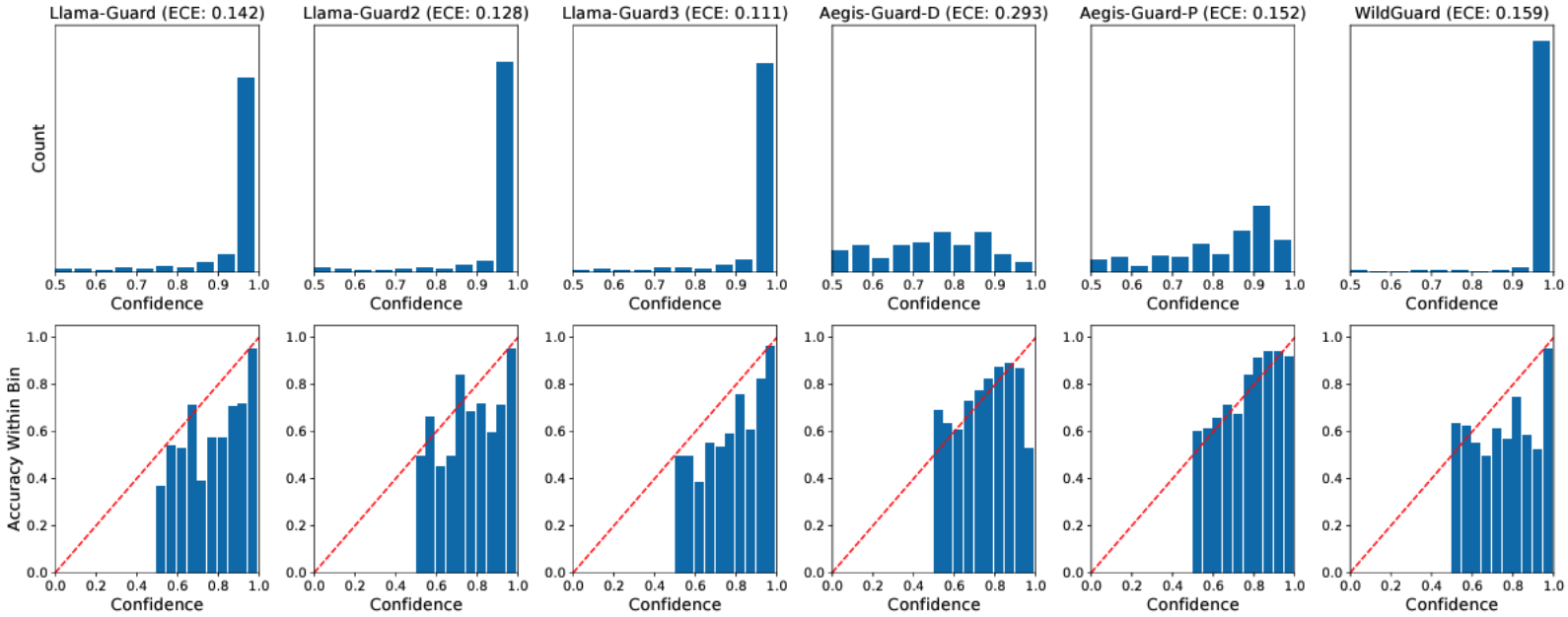
Finding 1

- Guardrail models tend to make **overconfident** predictions with high probability.
- **Prompt** Classification on WildGuardMix Prompt Test Set



Finding 1

- Guardrail models tend to make **overconfident** predictions with high probability.
- **Response** Classification on WildGuardMix Response Test Set



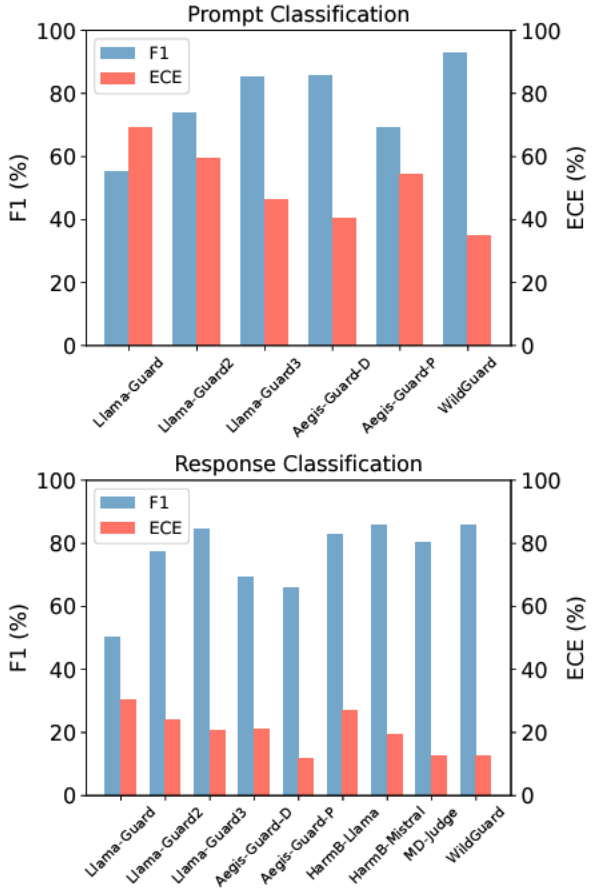
Finding 1

- Guardrail models tend to make **overconfident** predictions with high probability.
  - **WildGuard** stands out in prompt classification
  - **MD-Judge** stands out in response classification

Model	Prompt Classification								Response Classification				
	OAI	ToxiC	SimpST	Aegis	XST	HarmB	WildGT	Avg.	BeaverT	S-RLHF	HarmB	WildGT	Avg.
Llama-Guard	9.0	11.0	32.6	29.6	20.5	68.1	28.6	28.5	29.1	24.4	23.2	14.2	22.7
Llama-Guard2	13.7	15.9	26.5	34.7	12.2	30.3	26.8	22.9	29.9	25.2	24.9	12.8	23.2
Llama-Guard3	13.9	14.5	7.6	33.4	13.9	8.8	20.5	16.1	32.3	25.3	27.2	11.1	24.0
Aegis-Guard-D	30.2	20.5	9.7	16.4	23.5	50.5	8.0	22.7	18.1	30.9	26.1	29.3	26.1
Aegis-Guard-P	15.7	8.2	16.5	22.6	18.6	59.4	12.6	22.0	18.6	25.6	18.0	15.2	19.4
HarmB-Llama	-	-	-	-	-	-	-	-	24.9	19.4	15.1	52.5	28.0
HarmB-Mistral	-	-	-	-	-	-	-	-	18.1	14.5	13.1	25.6	17.8
MD-Judge	-	-	-	-	-	-	-	-	10.9	9.4	17.7	7.7	**11.4**
WildGuard	33.8	19.8	4.4	12.0	5.0	6.3	19.7	**14.4**	23.2	23.3	12.8	15.9	18.8

Finding 2

- Dataset: Harmbench
- Miscalibration in **prompt** classification is more pronounced than in **response** classification under **jailbreak attacks**.
- SOTA WildGuard achieves F1 score of 92.8% , but the ECE score remains 34.9%
- Larger shift in adversarial prompt distribution than LLM response distribution



Finding 3

- **Inconsistent reliability** when classifying outputs from different response LLMs
  - 10 different response LLMs
- **Limitation** of training data for guardrail models

Guard Model	Metric	Response Model									
		Baichuan2	Qwen	Solar	Llama2	Vicuna	Orca2	Koala	OpenChat	Starling	Zephyr
Llama-Guard	F1	57.8	66.7	54.2	44.4	64.0	62.7	74.6	60.6	66.7	72.7
	ECE	26.9	23.0	49.4	10.5	28.0	26.4	27.4	40.3	46.3	38.5
Llama-Guard2	F1	77.8	88.9	82.8	71.4	72.1	80.6	78.4	70.0	82.0	78.4
	ECE	18.2	5.8	27.1	7.9	28.4	25.2	30.4	28.5	37.0	39.4
Llama-Guard3	F1	73.8	82.4	84.1	60.0	83.3	82.2	77.5	76.4	91.2	87.7
	ECE	33.7	17.1	31.0	27.4	20.5	27.3	36.5	34.2	27.6	23.1
Aegis-Guard-D	F1	60.3	66.7	71.2	31.2	63.3	65.8	69.9	78.3	84.4	89.3
	ECE	35.5	27.1	22.2	40.8	34.0	33.9	31.3	30.9	27.8	30.9
Aegis-Guard-P	F1	57.6	66.7	67.9	33.3	56.3	72.7	72.2	76.2	83.3	80.8
	ECE	22.8	17.4	28.3	23.6	26.2	28.2	32.1	35.5	36.3	25.7
HarmB-Llama	F1	89.7	100.0	90.6	70.6	90.9	86.2	88.9	89.4	90.9	94.5
	ECE	17.7	6.4	25.0	23.5	16.0	19.5	26.7	23.2	23.3	20.1
HarmB-Mistral	F1	84.4	100.0	87.5	80.0	92.3	84.8	92.8	90.9	89.2	94.5
	ECE	28.0	3.0	30.1	12.8	16.6	17.5	16.0	14.9	27.7	19.9
MD-Judge	F1	75.4	79.1	77.2	55.6	74.2	76.9	75.3	76.6	87.5	92.6
	ECE	22.4	14.4	19.3	24.1	19.9	16.7	26.2	25.5	26.0	17.9
WildGuard	F1	82.0	91.3	88.5	80.0	89.9	84.8	81.6	88.9	92.5	94.5
	ECE	22.1	9.2	15.5	17.0	11.2	20.1	37.3	25.4	18.6	21.3

Improve Confidence Calibration

- Temperature Scaling (TS)
  - **Idea**: Apply **temperature (T) on logits** to smooth or sharp output distribution
  - **Limitation**: Require validation set
- Contextual Calibration (CC)
  - **Idea**: Estimate test-time contextual bias via **content-free token**, such as space token
  - **Limitation**: **inaccurate** to estimate the bias using content-free token
- Batch Calibration (BC)
  - **Idea**: Estimate test-time contextual bias via **a batch of unlabeled samples**
  - **Limitation**: need validation set to adjust proper **batch size**

Improve Confidence Calibration

Model	Prompt Classification								Response Classification				
	OAI	ToxiC	SimpST	Aegis	XST	HarmB	WildGT	Avg.	BeaverT	S-RLHF	HarmB	WildGT	Avg.
Llama-Guard	9.0	11.0	32.6	29.6	20.5	68.1	28.6	28.5	29.1	24.4	23.2	14.2	22.7
+ TS	12.0	11.3	31.9	26.8	9.4	66.7	26.0	26.3	27.4	21.6	14.5	14.0	19.4
+ CC	14.8	7.4	26.3	22.0	23.9	65.1	20.9	**25.8**	25.4	21.8	20.2	8.9	**19.1**
+ BC	12.3	12.1	43.2	27.2	21.0	67.9	19.7	29.1	26.6	22.5	20.6	12.4	20.5
Llama-Guard2	13.7	15.9	26.5	34.7	12.2	30.3	26.8	22.9	29.9	25.2	24.9	12.8	23.2
+ TS	13.2	15.8	26.0	33.6	11.1	29.4	26.0	22.2	29.8	24.5	14.1	13.6	**20.5**
+ CC	39.4	22.8	15.0	18.8	13.7	14.8	15.3	**20.0**	24.3	28.9	34.8	32.8	30.2
+ BC	15.2	16.6	30.6	34.2	12.0	36.3	23.8	24.1	29.5	25.2	24.8	14.7	23.6
Llama-Guard3	13.9	14.5	7.6	33.4	13.9	8.8	20.5	16.1	32.3	25.3	27.2	11.1	24.0

Takeaway:

- In general**, CC** proves more effective for **prompt** classification and **TS** benefits **response** classification more.
- **In-domain validation sets** help determine a better temperature/batch size and yield better calibration results.
- No single post-hoc calibration method fully resolves miscalibration.

WildGuard	33.8	19.8	4.4	12.0	5.0	6.3	19.7	14.4	23.2	23.3	12.8	15.9	18.8
+ TS	32.4	19.1	5.7	9.1	4.2	8.2	19.3	**14.0**	23.8	22.3	10.5	16.5	**18.3**
+ CC	58.7	39.0	0.2	26.5	25.5	0.1	18.6	24.1	22.8	27.9	16.2	16.1	20.8
+ BC	33.6	23.8	25.2	12.7	3.8	30.6	19.5	21.3	23.1	22.2	12.6	16.3	18.6

Thanks

- Check our paper for more details

Paper

- We release the evaluation tool and advocate for incorporating evaluation of confidence calibration when releasing future LLM-based guardrail models.

GitHub