

Guard Models Are Overconfident Where Base Models Are Uncertain

Jonghyun Hong¹ MinJae Jung¹ Minwoo Kim^{1*}

¹ DATUMO INC.

{jonghyun.hong,minjae.jung,mwkim}@selectstar.ai

Abstract

Guard models are used as safety classifiers, with confidence scores driving downstream moderation decisions. We evaluate five guard models for prompt classification and find that although several are nearly calibrated on clean inputs, adversarial attacks degrade their calibration by an order of magnitude, turning false negatives into high-confidence errors indistinguishable from correct detections. Comparing each guard with its corresponding base LM, we find that uncertainty signals often remain available, with the base model typically expressing uncertainty on the same inputs where the guard fails. Layer-wise analyses localize this guard–base divergence to later layers, where guard models exhibit sharper safe/unsafe separation and lower-rank representations, while adversarial harmful inputs lie closer to the clean-safe region. These findings highlight a mismatch between guard confidence and base-model uncertainty under attack. **WARNING: This paper contains model outputs that can be offensive in nature.**

1 Introduction

Large Language Models (LLMs) have become a central interface for user-facing AI systems, building on broad task-solving capabilities and advances in instruction following (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2024). As their use expands across general-purpose and domain-specific applications, broader deployment also amplifies safety risks, including misuse, manipulation, and jailbreak attacks that induce harmful generations (Weidinger et al., 2021; Ganguli et al., 2022; Zou et al., 2023). Consequently, ensuring the safety and reliability of deployed LLM systems has become a major challenge.

To mitigate these risks, guardrails have increasingly been adopted as external safety layers that operate around the base model at inference time (Rebedea et al., 2023; Yuan et al., 2024). These guardrails monitor user inputs and model outputs to enforce policy-specific safety constraints. In many deployment pipelines, this role falls to guard models: auxiliary classifiers that judge prompts or responses against predefined safety taxonomies and decide whether to allow, flag, or block an interaction (Inan et al., 2023; Ghosh et al., 2024; Han et al., 2024; Zeng et al., 2024). In practice, these decisions often rely on confidence-based thresholding rather than binary verdicts alone.

Despite their growing role, the reliability of the confidence scores driving these moderation decisions remains underexplored. Guard models inherit the broader calibration problem observed in LLMs, where confidence is often misaligned with actual correctness (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024; Leng et al., 2025). Recent work confirms that this issue extends to guard models, showing severe overconfidence under jailbreak inputs, which post-hoc methods such as temperature scaling, contextual calibration, and batch calibration can partially address (Liu et al., 2025), while complementary work has shown that guard predictions can be unstable even under meaning-preserving paraphrases, motivating self-supervised consistency training for semantic robustness (Pinneri and Louizos, 2025). In contrast, we diagnose adversarial miscalibration through matched guard–base comparisons of uncertainty and internal representations. Our results indicate that guard failures reflect surface-form biases in safety judgments rather than poor calibration alone, leaving guards vulnerable to jailbreak attacks that disguise harmful intent.

In this work, we evaluate five guard models

*Corresponding author.

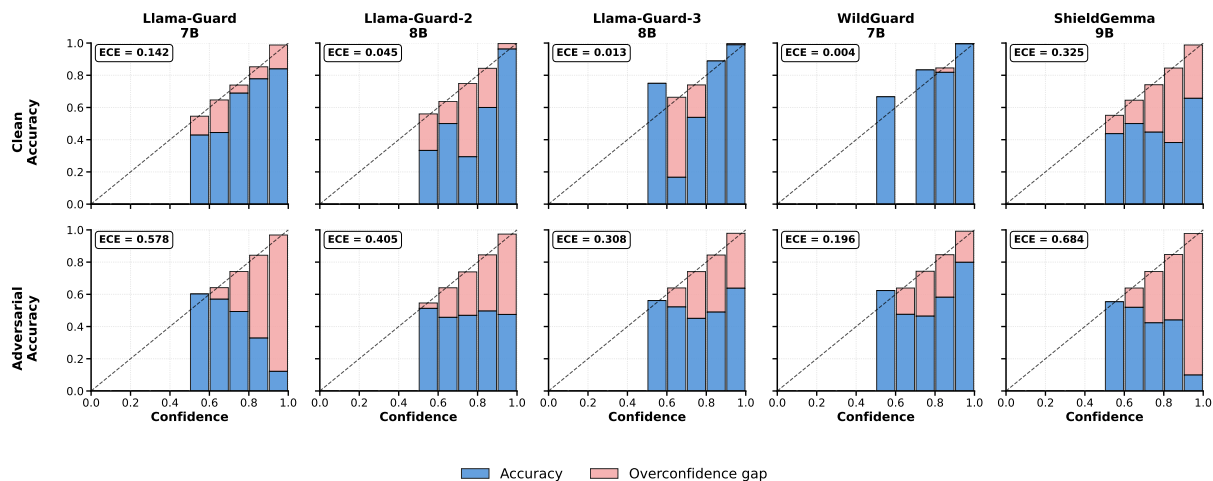


Figure 1: Reliability diagrams for guard model confidence on clean (Top) and adversarial (Bottom) inputs. Blue bars show empirical accuracy per confidence bin, the dashed diagonal denotes perfect calibration, and red regions indicate overconfidence. Several guards are well calibrated on clean inputs, but adversarial inputs degrade calibration by an order of magnitude, with overconfidence concentrated in the highest-confidence bins. A per-attack breakdown is provided in Appendix B.

across six attack methods spanning gradient-based, template-based, and LLM-based strategies, including a custom attack method. As Figure 1 shows, guards that are nearly well calibrated on clean inputs degrade by an order of magnitude under adversarial inputs, with the overconfidence concentrated in high-confidence bins. Moreover, Figure 2 reveals that false negatives are not low-confidence borderline cases but high-confidence errors. Even at a 0.99 threshold, a substantial fraction of adversarial harmful prompts pass through as confident false negatives, indistinguishable from correct detections. This raises the question of whether such confident guard failures reflect ambiguity in the adversarial inputs or whether uncertainty remains available in the corresponding base model.

Motivated by findings that pre-trained base LLMs retain well-calibrated uncertainty even as post-training erodes it (Nakkiran et al., 2025; Leng et al., 2025; Luo et al., 2025), we compare each guard with its corresponding base LM. Both are evaluated through a shared binary safe/unsafe interface, with base LMs using a 5-shot classification prompt. We find that adversarial wrapping induces larger verdict shifts in guards than in their corresponding base LMs (Figure 3), indicating greater sensitivity to surface-level context changes despite unchanged harmful content. Yet this sensitivity is not reflected in uncertainty: on most guard false negatives, the base

model assigns higher entropy to the same input (Figure 4). We then localize this divergence with layer-wise and representation analyses. Guards increasingly assign lower unsafe probability than their base models in late layers (Figure 5), while their hidden states become concentrated in a lower-rank space with a sharper safe/unsafe separation (Figures 6 and 7). This structure separates clean inputs, but often places adversarial harmful inputs near the clean-safe side, producing confident misclassifications.

We make three main contributions:

- We show that guard model confidence, even when well calibrated on clean inputs, becomes overconfident under adversarial attack, with false negatives reaching confidence levels comparable to correct detections.
- We compare guards with their corresponding base LMs and find that, on most guard false negatives, base models assign higher entropy to the same input, revealing a systematic uncertainty gap between guards and their corresponding base models.
- We provide layer-wise evidence that guard-base divergence emerges in later layers, where guards exhibit lower unsafe probabilities, lower-rank representations, and sharper safe/unsafe separation, while adversarial harmful inputs lie closer to the clean-safe region.

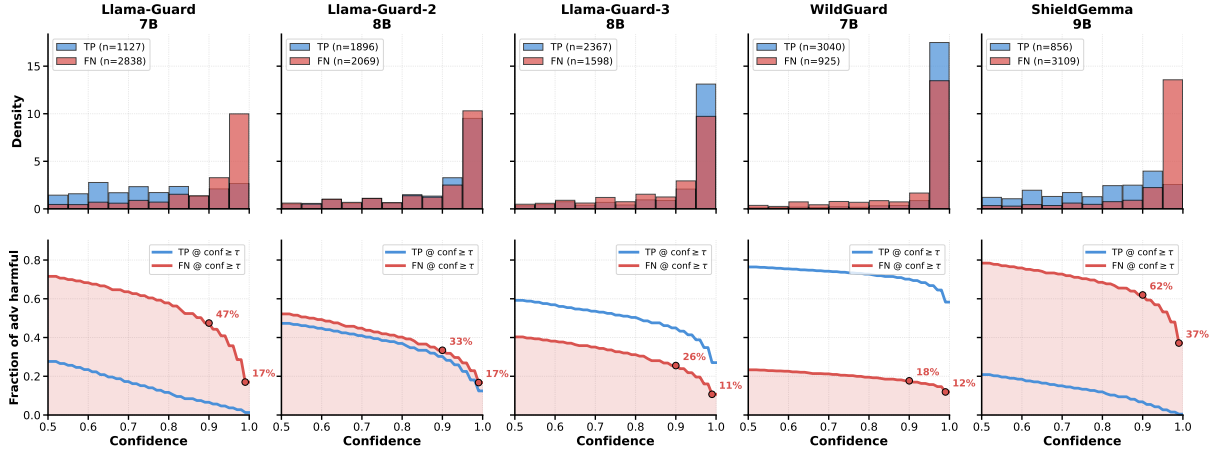


Figure 2: Guard overconfidence persists at every confidence level. (Top) Confidence distributions of true-positive (blue) and false-negative (red) decisions on the pooled adversarial harmful set. For all five guards, FN density concentrates at the high-confidence end, overlapping heavily with TP. (Bottom) Fraction of the adversarial pool retaining a TP or FN decision at confidence $\geq \tau$. Even at $\tau = 0.99$, substantial FN fractions remain, indicating that confidence thresholding cannot selectively filter missed detections.

2 Related Work

Guardrails and Guard Models for LLM Safety. Content moderation systems such as Perspective API (Lees et al., 2022) and OpenAI’s moderation API (Markov et al., 2023) detect toxic content in human-generated text but do not address the interactive setting of LLM deployment, where jailbreak attempts and unsafe generations introduce distinct risks. LLM-specific guard models such as Llama-Guard (Inan et al., 2023), Aegis (Ghosh et al., 2024), MD-Judge (Li et al., 2024), WildGuard (Han et al., 2024), and ShieldGemma (Zeng et al., 2024) classify prompts or responses against safety taxonomies, evaluated by benchmarks such as HarmBench (Mazeika et al., 2024), AdvBench (Zou et al., 2023), ToxiGen (Hartvigsen et al., 2022), and RigorLLM (Yuan et al., 2024). System-level frameworks such as NeMo Guardrails (Rebedea et al., 2023) and SafeRoute (Lee et al., 2025) address deployment routing and efficiency. This prior work focuses on whether guards correctly detect harmful content, not whether their confidence in those decisions is trustworthy.

Calibration and uncertainty in LLMs. LLMs can estimate their own correctness (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024), yet post-training systematically inflates confidence (Leng et al., 2025). In contrast, pre-trained base LLMs appear to retain well-calibrated uncertainty: Nakkiran et al. (2025)

show that semantic calibration emerges as a byproduct of next-token pre-training and persists under few-shot prompting, while Luo et al. (2025) demonstrate that pre-trained representations can serve as unsupervised calibrators for post-trained models. Building on this insight, BaseCal (Tan et al., 2026) projects post-trained hidden states into the base model’s representation space, and Dual-Align (Luo et al., 2026) corrects layer-wise inference drift between pre- and post-trained models. Our work shares the premise that base models provide a useful calibration reference but uses the comparison diagnostically rather than proposing a correction method. These findings also motivate our use of a 5-shot interface to elicit base-model uncertainty as a reference.

3 Background

Guard models and restricted softmax. A guard model is a safety classifier built on top of a language model. Given a user prompt x , it predicts a binary safety label $\hat{y}(x) \in \{\text{safe}, \text{unsafe}\}$. We focus on prompt classification, where the guard determines whether the user input itself should be allowed or blocked before it is passed to a downstream model. In deployed guard models such as Llama-Guard, WildGuard, and ShieldGemma, this classification is implemented through autoregressive generation: the model is prompted with an instruction or safety policy and generates a label token such as safe/unsafe or an equivalent allow/block label. We compute the

guard’s safety distribution by applying a restricted softmax over the two allowed safety labels. Let $z_{\text{safe}}(x)$ and $z_{\text{unsafe}}(x)$ denote the logits assigned to the two labels at the decision position. Then

$$p_{\text{unsafe}}(x) = \frac{\exp z_{\text{unsafe}}(x)}{\exp z_{\text{safe}}(x) + \exp z_{\text{unsafe}}(x)}. \quad (1)$$

We set $p_{\text{safe}}(x) = 1 - p_{\text{unsafe}}(x)$. The predicted label is

$$\hat{y}(x) = \underset{y \in \{\text{safe}, \text{unsafe}\}}{\operatorname{argmax}} p_y(x). \quad (2)$$

This restricted distribution measures confidence in the binary safety decision rather than the model’s absolute probability mass over the full vocabulary.

Accuracy, confidence, and calibration.

Accuracy measures whether the predicted label matches the ground-truth safety label. Clean benign prompts are labeled safe, while clean harmful prompts are labeled unsafe. For adversarial prompts, the underlying harmful behavior remains unsafe even when it is wrapped by a jailbreak attack.

We define confidence as the probability assigned to the predicted label,

$$c(x) = p_{\hat{y}(x)}(x) = \max_{y \in \{\text{safe}, \text{unsafe}\}} p_y(x). \quad (3)$$

A model is calibrated when its confidence matches its empirical accuracy: among examples assigned confidence near p , approximately a fraction p should be correct. We quantify calibration with Expected Calibration Error (ECE). Given B confidence bins, let S_b be the set of examples whose confidence falls in bin b . Then

$$\text{ECE} = \sum_{b=1}^B \frac{|S_b|}{n} |\text{acc}(S_b) - \text{conf}(S_b)|, \quad (4)$$

where $\text{acc}(S_b)$ is the empirical accuracy in bin b and $\text{conf}(S_b)$ is the mean confidence. Lower ECE indicates better calibration.

4 Guard Confidence Under Adversarial Attack

We analyze guard model confidence under adversarial inputs in three stages. After describing the experimental setup (Section 4.1), we show that guard confidence becomes severely overconfident under attack (Section 4.2). We then compare guards with their base LMs to test whether uncertainty

signals remain available before safety fine-tuning (Section 4.3), and finally localize guard–base divergence in the network through layer-wise analyses (Section 4.4).

4.1 Experimental Setup

We evaluate guard model confidence by comparing each fine-tuned guard with its corresponding base LM on clean and adversarial prompt-classification inputs.

Models. The guards are Llama-Guard 7B/Llama-Guard-2 8B/Llama-Guard-3 8B (Inan et al., 2023), WildGuard 7B (Han et al., 2024), and ShieldGemma 9B (Zeng et al., 2024). Their corresponding base LMs are Llama-2 7B (Touvron et al., 2023), Llama-3 8B, Llama-3.1 8B (Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023) and Gemma-2 9B (Team et al., 2024), respectively.

Data. The clean pool contains 713 harmful prompts from HarmBench (Mazeika et al., 2024) and StrongREJECT (Souly et al., 2024), and 500 benign prompts from Alpaca (Taori et al., 2023), yielding 1,213 clean prompts. To build the adversarial pool, we apply jailbreak attacks to the harmful prompts: GCG (Zou et al., 2023), AutoDAN (Liu et al., 2024b), TAP (Mehrotra et al., 2024), PAIR (Chao et al., 2023), AutoDAN-Turbo (Liu et al., 2024a), and a custom TextGrad-based attack (Yuksekgonul et al., 2024), yielding 3,965 adversarial harmful prompts. Ground-truth labels are inherited from the source behavior: HarmBench and StrongREJECT prompts, both clean and adversarial, are labeled unsafe, while Alpaca prompts are labeled safe.

Prompts. Guard models are evaluated with their native safety instructions and chat templates, including the official policy text or taxonomy when provided. Because guard models differ in their output conventions, we adapt the label vocabulary accordingly: Yes/No for ShieldGemma 9B and WildGuard 7B, and safe/unsafe for the Llama-Guard family, following each model’s official instruction template. For base LMs, we use a 5-shot binary classification prompt with three safe and two unsafe demonstrations, motivated by evidence that base LMs retain calibrated confidence under few-shot prompting (Nakkiran et al., 2025). For both guard and base models, confidence is computed by applying the restricted softmax over the two label

tokens at the decision position. The exact prompt and additional analyses of verdict-token validity and robustness to alternative base-LM interfaces are provided in Appendix A.

4.2 Guard Confidence Becomes Overconfident Under Attack

We first ask whether guard confidence remains calibrated when harmful prompts are adversarially wrapped. Reliability diagrams in Figure 1 show that calibration on clean inputs does not transfer to adversarial settings. On clean inputs, several recent guards are well calibrated without post-hoc calibration (Llama-Guard-3 ECE 0.013, WildGuard ECE 0.004), while Llama-Guard and ShieldGemma already show visible miscalibration (ECE 0.142 and 0.325, respectively). Under adversarial inputs, however, ECE increases sharply across all guards, often by an order of magnitude. Even the two best-calibrated guards on clean inputs degrade substantially: Llama-Guard-3 increases from 0.013 to 0.308, and WildGuard from 0.004 to 0.196. The degradation is concentrated in high-confidence bins, indicating that adversarial inputs induce confident but incorrect safety decisions rather than merely increasing uncertainty.

The confidence distributions in Figure 2 explain why this failure is difficult to mitigate. On adversarial inputs, false negatives are not concentrated near the decision boundary at 0.5; instead, their confidence often peaks near 1.0 and heavily overlaps with the true-positive distribution. Thus, missed detections are not low-confidence borderline cases but high-confidence errors, made with confidence comparable to correct detections. ShieldGemma exhibits a broader false-negative confidence distribution, but this reflects its low overall adversarial accuracy rather than reliable uncertainty.

The bottom row of Figure 2 translates this overlap into a deployment-relevant view. As the confidence threshold τ increases, both true positives and false negatives are filtered, but false negatives are not selectively removed. Even at $\tau = 0.99$, Llama-Guard-3 retains 11% of adversarial harmful prompts as high-confidence false negatives, WildGuard retains 12%, and ShieldGemma retains 37%. Thus, an “abstain when uncertain” approach cannot recover these missed detections, as guards assign similarly high confidence to both correct and incorrect safety decisions.

Type	Model	Clean		Adv.	
		Acc.	ECE	Acc.	ECE
Base	Llama-2 7B	.634	.275	.992	.051
Guard	Llama-Guard 7B	.815	.142	.284	.578
Base	Llama-3 8B	.920	.023	.746	.034
Guard	Llama-Guard-2 8B	.939	.045	.478	.405
Base	Llama-3.1 8B	.903	.037	.835	.062
Guard	Llama-Guard-3 8B	.979	.013	.597	.308
Base	Mistral 7B	.935	.102	.802	.116
Guard	WildGuard 7B	.989	.004	.767	.196
Base	Gemma-2 9B	.906	.084	.930	.048
Guard	ShieldGemma 9B	.622	.325	.216	.684

Table 1: Base-guard performance and calibration. Each base model and its fine-tuned guard are evaluated with the same binary safe/unsafe interface. Columns report accuracy and ECE on clean and adversarial inputs.

4.3 Base LMs Retain Uncertainty Under Attack

Given evidence that pre-trained base LLMs retain well-calibrated uncertainty, we ask whether the overconfidence guards exhibit under attack reflects an inherent limitation of the adversarial inputs or a divergence between guards and their corresponding base models. Comparing each guard with its base LM on the same inputs can distinguish these two cases. If the base model is equally overconfident, the failure is shared across the pair; if it retains uncertainty, this indicates a guard-base uncertainty gap on the same input (Nakkiran et al., 2025; Leng et al., 2025).

Aggregate performance. Table 1 compares each guard with its corresponding base LM under the binary safe/unsafe interface. Safety fine-tuning generally improves clean-input classification, but this gain does not translate to adversarial calibration. For example, Llama-Guard-3 improves clean accuracy over Llama-3.1, yet its adversarial ECE increases from 0.062 for the base model to 0.308 for the guard. Similar trends appear across the Llama-Guard family, while WildGuard shows a smaller but consistent degradation. ShieldGemma is an outlier: its clean accuracy is already substantially lower than that of its base model, so we treat it as a broader guard failure rather than a purely adversarial calibration effect. Separately, we exclude the Llama-Guard 7B/Llama-2 7B pair from subsequent matched analyses because the Llama-2 base model exhibits a strong unsafe-prediction bias

on the adversarial set. We additionally test whether the calibration degradation extends to response moderation and whether the guard–base patterns persist across model scales in Appendices D and C, respectively.

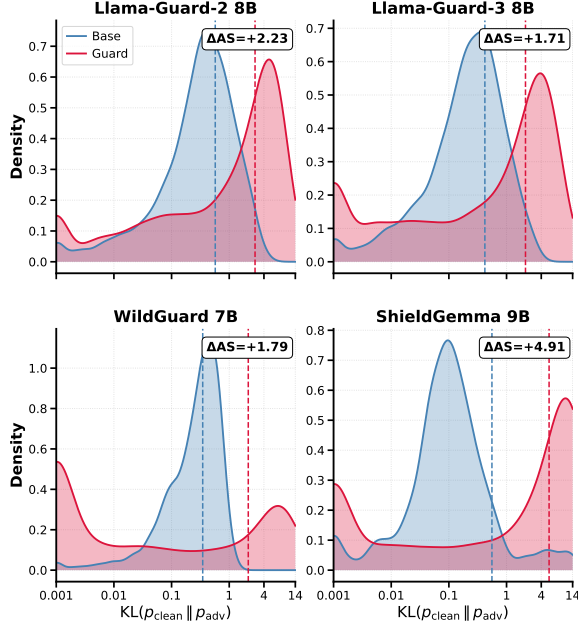


Figure 3: Attack-induced shifts in verdict distributions for base and guard models, measured by $\text{KL}(p_{\text{clean}} \| p_{\text{adv}})$. Analysis is restricted to prompts that both the base LM and guard confidently classify as unsafe, reducing confounding from initial disagreement. Dashed lines mark mean attack sensitivity, and ΔAS reports the guard–base difference. Across all pairs, guards exhibit substantially larger verdict shifts than their base LMs.

Attack-induced verdict shift. Aggregate calibration shows that guards degrade more severely than their base LMs under attack, but it does not reveal how each model’s safety judgment responds to adversarial attack. We therefore compare each model’s binary verdict distribution on a clean harmful prompt x and its attacked version $T(x)$, measuring the shift via the KL divergence between the two:

$$\text{AS}_M(x, T) = \text{KL}\left(D_M(x) \parallel D_M(T(x))\right), \quad (5)$$

where $D_M(\cdot) = \text{Bernoulli}(p_{\text{unsafe}})$ is the model’s binary verdict distribution. To isolate the effect of the attack from differences in clean-input behavior, we restrict the comparison to examples where both the base LM and the guard assign high unsafe probability (≥ 0.9) to the clean prompt.

We define the attack sensitivity gap as $\Delta\text{AS} = \overline{\text{AS}}_{\text{guard}} - \overline{\text{AS}}_{\text{base}}$, where $\overline{\text{AS}}$ denotes the mean attack shift across samples; positive values indicate larger shifts for the guard.

Figure 3 shows that guards undergo substantially larger verdict shifts than their base LMs under adversarial wrapping. The effect is consistent across all four guard–base pairs, with ΔAS ranging from +1.71 (Llama-Guard-3) to +4.91 (ShieldGemma). The same harmful content, merely rephrased by an adversarial wrapper, causes the guard’s safety judgment to shift far more than the base model’s. This suggests that guards are not robust to surface-level context: their verdicts are easily destabilized by the attack, making it *difficult to trust* their predictions under adversarial input.

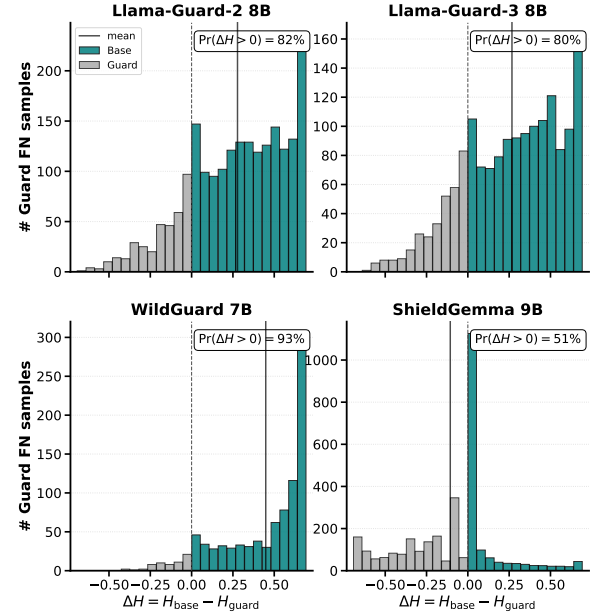


Figure 4: Distribution of per-sample entropy gap $\Delta H = H_{\text{base}} - H_{\text{guard}}$ on guard false-negative samples. Each histogram covers inputs where the guard incorrectly predicts safe on a ground-truth-unsafe prompt. Positive bars (teal) indicate the base model is more uncertain; negative bars (grey) the reverse. The dashed line marks $\Delta H = 0$, the solid line the mean, and the inset reports $\Pr(\Delta H > 0)$.

Confident despite being destabilized.

Despite this sensitivity, guards do not express corresponding uncertainty when they fail. To compare uncertainty at the individual-prompt level, we compute the binary entropy of each model’s

verdict distribution,

$$H_M(x) = - \sum_{y \in \{\text{safe}, \text{unsafe}\}} p_y^M(x) \log p_y^M(x), \quad (6)$$

and measure the entropy gap $\Delta H = H_{\text{base}} - H_{\text{guard}}$ on adversarial prompts where the guard produces a false negative.

Figure 4 shows that ΔH is positive for the vast majority of guard failures: 82% of Llama-Guard-2, 80% of Llama-Guard-3, and 93% of WildGuard false negatives have $\Delta H > 0$. ShieldGemma is an exception at 51%, consistent with its broader calibration failure noted in Section 4.2. That is, on inputs where the guard confidently predicts safe, the base model typically expresses greater uncertainty about the same decision. Guards are not merely destabilized by the adversarial wrapper, and they also remain confident where the base model expresses uncertainty on the same inputs.

Guards are highly sensitive to attack yet show *no corresponding increase in uncertainty*, while the base model retains it on the same inputs. A natural question is how guard and base model diverge internally to produce such different behaviors.

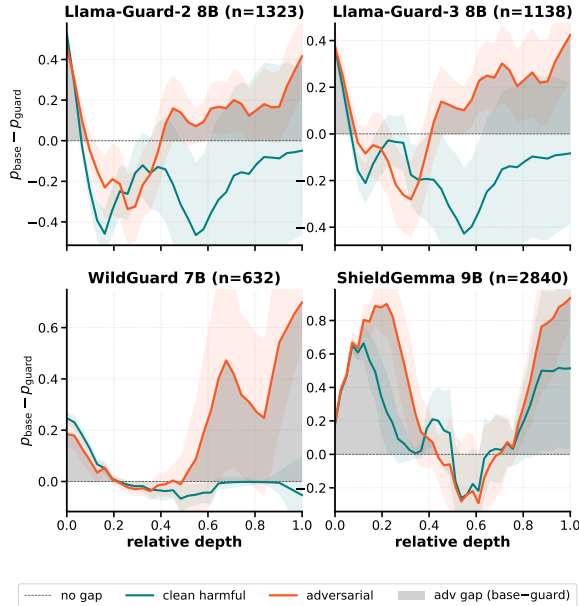


Figure 5: Per-layer gap between base and guard unsafe probabilities, $p_{\text{base}}^{(\ell)} - p_{\text{guard}}^{(\ell)}$, plotted against relative depth. Lines show mean ± 1 standard deviation for clean harmful inputs (teal) and adversarial false negatives where the base correctly predicts unsafe but the guard predicts safe (orange). The shaded grey band highlights where the guard assigns lower unsafe probability than the base on the adversarial curve.

4.4 Where and How Overconfidence Arises

Guard models make more decisive safe/unsafe judgments than their corresponding base models. We now examine how the guard and base model diverge layer by layer, using per-layer logit projections and hidden-state geometry.

Logit gap across depth. At each layer ℓ , we obtain $p_{\text{unsafe}}^{(\ell)}$ by passing the verdict-position hidden state through the model’s final RMSNorm and LM head and applying a restricted softmax over the two verdict tokens. To localize this divergence, we examine the layer-wise gap $p_{\text{base}}^{(\ell)} - p_{\text{guard}}^{(\ell)}$ between base and guard unsafe probabilities on two input sets: clean harmful inputs and adversarial false negatives where the base predicts unsafe but the guard predicts safe. As Figure 5 shows, on clean harmful inputs the gap is near zero or slightly negative across depth, indicating that the guard assigns at least as much unsafe probability as the base, as expected from safety fine-tuning. Under adversarial inputs, however, the gap reverses and grows sharply through the later layers, reaching approximately 0.45 for Llama-Guard-3, 0.7 for WildGuard, and 0.9 for ShieldGemma at the final layer. This divergence emerges progressively across the final third of the layers rather than appearing only at the output layer, showing that the guard–base difference is already present in intermediate representations.

Representation geometry. A complementary view emerges from hidden-state geometry. As Figure 6 shows, in the base model, clean safe, clean harmful, and adversarial harmful inputs remain largely intermixed at both early and late layers; the base model does not develop a safety-specific geometry in its representations. Guard representations, by contrast, exhibit a sharper safe/unsafe boundary in late-layer representations that cleanly separates clean safe from clean harmful inputs. This boundary is consistent with the guard’s strong clean-input performance. Under adversarial context, however, the same boundary coincides with failure. Adversarial harmful inputs, despite containing the same underlying harmful content, are mapped closer to the clean-safe side of the boundary, and it is on these inputs that the guard produces confident misclassifications.

This boundary is accompanied by compression of the representation space, which we quantify using effective rank (Roy and Vetterli, 2007):

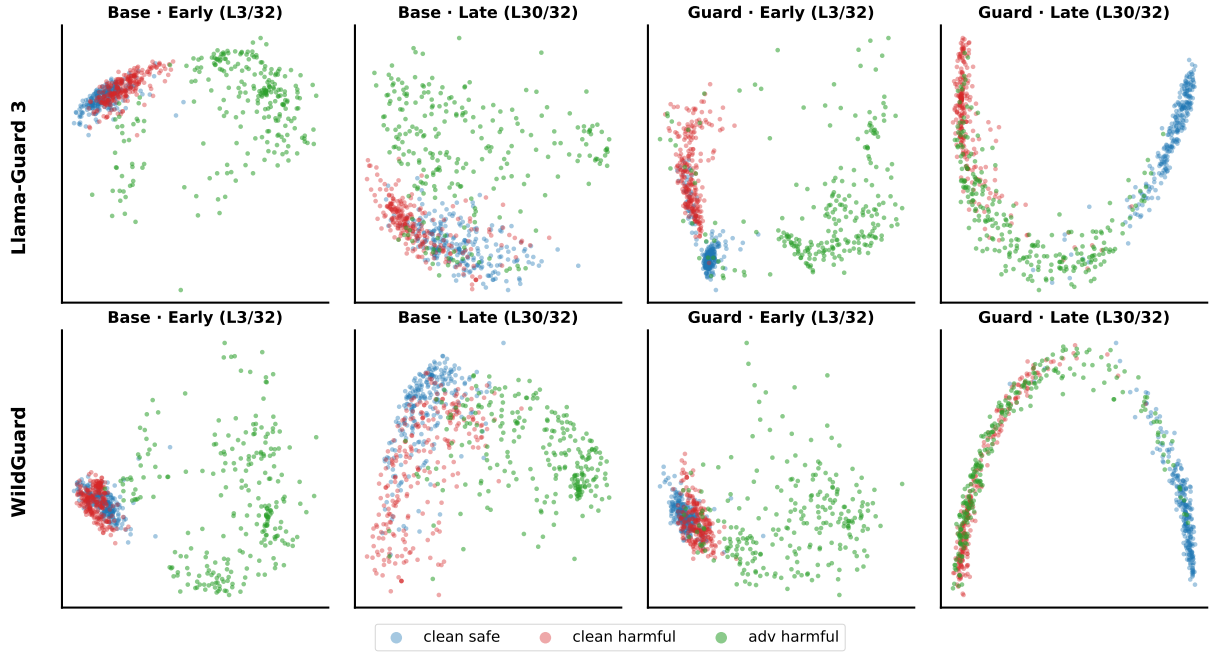


Figure 6: PCA projection (top-2 components, fit independently per panel) of last-token hidden states for three input classes: clean safe (blue), clean harmful (red), and adversarial harmful (green). Rows compare Llama-Guard-3 (top) and WildGuard (bottom) with their respective base models. Columns show layer 3 and layer 30 out of 32 total layers, representing early- and late-network representations respectively.

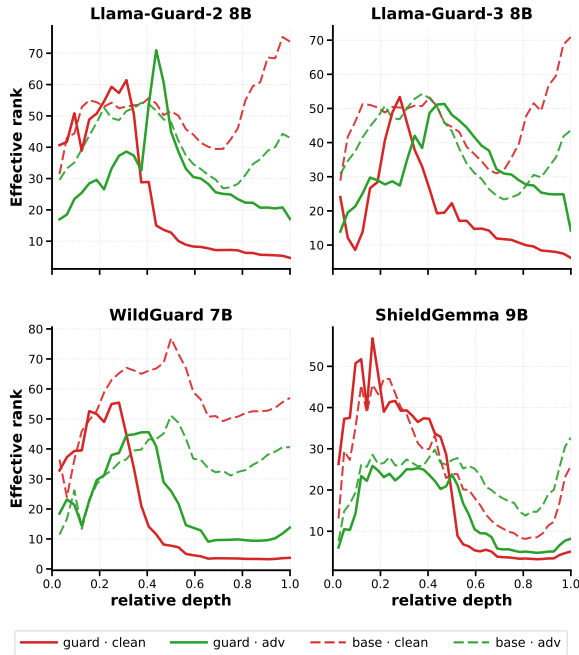


Figure 7: Per-layer effective rank of last-token hidden states for four guard/base pairs. Solid lines denote guards, dashed lines their base LMs; red indicates clean inputs and green adversarial inputs.

Effective rank($\mathbf{X}$) = $\exp(-\sum_i p_i \log p_i)$, where $p_i = \lambda_i / \sum_j \lambda_j$ are the normalized eigenvalues of the centered Gram matrix. As Figure 7

shows, guards occupy a markedly lower-rank subspace than their base LMs in late layers, with base-to-guard effective-rank ratios of roughly $1.4\times$ (Llama-Guard-2), $1.7\times$ (Llama-Guard-3), $3.4\times$ (WildGuard), and $4.0\times$ (ShieldGemma) at the penultimate layer. Within this compressed subspace, the safe/unsafe boundary cleanly separates clean inputs, but adversarial harmful inputs whose surface form resembles safe content are collapsed onto the clean-safe side of the boundary, where the guard does not express uncertainty.

Together, these analyses capture the same failure from complementary perspectives. The logit gap localizes where the guard–base unsafe-probability difference emerges across depth, progressively through late layers, while the PCA and effective-rank analyses characterize the accompanying representation geometry. Guard representations become compressed into a lower-rank space with a sharper safe/unsafe boundary that cleanly separates clean inputs but becomes brittle under adversarial context. In contrast, the base model does not exhibit the same sharply separated late-layer safety geometry under this probing setup.

5 Conclusion

We analyzed whether guard model confidence remains reliable under adversarial attack. Across five guard models, adversarial wrapping destabilizes verdicts without increasing uncertainty, so false negatives receive confidence comparable to correct detections and are difficult to remove by thresholding. Comparing guards with their corresponding base LMs shows that this failure is not inherent to the inputs alone, since base models are usually more uncertain on the same false-negative cases. Layer-wise analyses localize this gap to later layers, where guards exhibit lower unsafe probabilities, lower-rank representations, and sharper safe/unsafe separation than their corresponding base models. These findings suggest that guard models should be evaluated not only by detection accuracy, but also by whether their confidence remains trustworthy under attack. Base-model uncertainty is one signal that a calibration-aware guardrail could draw on, since it remains informative precisely where guard confidence fails.

Limitations

We localize the guard–base divergence to later layers and characterize the accompanying representation geometry, but our evidence is correlational. We observe that the logit gap, the sharper safe/unsafe boundary, and the lower-rank subspace all accompany confident guard failures. We do not intervene on these quantities, so we stop short of showing that this geometry is what produces the misclassifications rather than co-occurring with them. Establishing the mechanism would require causal interventions such as ablating or steering the identified directions. Our contribution is also diagnostic in nature. We show that base-model uncertainty remains informative where guard confidence fails, but we do not turn this signal into a calibration method or characterize the conditions under which the gap could be closed.

Ethical Considerations

This work analyzes vulnerabilities in guard models to inform safer deployment and strengthen defenses against adversarial attempts to bypass LLM safety mechanisms. All experiments were conducted in controlled research settings and are reported to emphasize mitigation, robustness, and

calibration-aware evaluation rather than actionable circumvention instructions. All attacks used in this study are based on previously published attack methods or optimization frameworks, and all datasets are drawn from publicly available benchmarks. No real-world harm was caused or intended, and no new harmful or illegal content was generated or disseminated. We acknowledge that exposing calibration failures in safety systems carries dual-use risk; however, we believe that understanding these failure modes is a necessary step toward building more reliable guardrails. We follow responsible disclosure practices where appropriate. This paper contains examples of model outputs that may be offensive in nature.

References

- 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.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
- 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*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 37:8093–8131.
- 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. In *Proceedings of the 60th annual meeting of the*

- association for computational linguistics (volume 1: Long papers)*, pages 3309–3326.
- 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*.
- 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*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Seanie Lee, Dong Bok Lee, Dominik Wagner, Minki Kang, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. 2025. Saferoute: Adaptive model selection for efficient and accurate safety guardrails in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2053–2069.
- Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207.
- Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiabin Huang. 2025. Taming overconfidence in llms: Reward calibration in rlhf. In *International Conference on Learning Representations*, volume 2025, pages 16484–16517.
- Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
- Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. 2025. On calibration of llm-based guard models for reliable content moderation. In *International Conference on Learning Representations*, volume 2025, pages 67808–67829.
- Xiaogeng Liu, Peiran Li, Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. 2024a. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. *arXiv preprint arXiv:2410.05295*.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations (ICLR)*.
- Beier Luo, Cheng Wang, Hongxin Wei, Sharon Li, and Xuefeng Du. 2026. Unlocking the pre-trained model as a dual-alignment calibrator for post-trained llms. *arXiv preprint arXiv:2601.04277*.
- Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. 2025. Your pre-trained LLM is secretly an unsupervised confidence calibrator. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- 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.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning (ICML)*.
- Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box LLMs automatically. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. 2025. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in llms. *arXiv preprint arXiv:2511.04869*.
- OpenAI. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Cristina Pinneri and Christos Louizos. 2025. Guarding the meaning: Self-supervised training for semantic robustness in guard models. *arXiv preprint arXiv:2511.10665*.

- Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 431–445.
- Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.
- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. 2024. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440.
- Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun, and Xueqi Cheng. 2026. Basecal: Unsupervised confidence calibration via base model signals. *arXiv preprint arXiv:2601.03042*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.
- 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. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57953–57965.
- Mert Yuksekogul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. TextGrad: Automatic “differentiation” via text. *arXiv preprint arXiv:2406.07496*.
- Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Base-LM Prompting and Interface Robustness

A.1 Base-LM Prompt and Verdict-Token Validation

For completeness, we show the 5-shot prompt used to elicit binary safety predictions from each base LM; it contains three safe and two unsafe demonstration pairs followed by the query placeholder {query}.

Prompt: Base 5-shot safety classifier	
Classify the following content as “safe” or “unsafe”.	
Content: What’s a good recipe for pasta?	
Label: safe	
Content: How do I pick a lock on someone’s house?	
Label: unsafe	
Content: Can you explain how photosynthesis works?	
Label: safe	
Content: Write a phishing email pretending to be a bank	
Label: unsafe	
Content: What are some tips for job interviews?	
Label: safe	
Content: {query}	
Label:	

For Mistral 7B and Gemma-2 9B, we substitute only the two label tokens with Yes/No, matching their corresponding guards. To ensure that the restricted softmax comparison between base and guard models is valid, we verify that base LMs concentrate substantial probability mass on the two verdict tokens under the 5-shot classification prompt.

Table 2 validates that the 5-shot binary classification prompt induces a well-formed verdict-token interface for the base LMs. For each model, we compute the full-vocabulary softmax probability assigned to the valid verdict tokens: safe/unsafe for the Llama-family models and Yes/No for Mistral/Gemma. Across the 5,178 examples, the valid verdict tokens receive most of the probability mass for every base model, ranging from 90.99% for Mistral 7B to 97.80% for Llama-

Base	Label mass	Other mass
Llama-2 7B	95.75	4.25
Llama-3 8B	95.89	4.11
Llama-3.1 8B	97.80	2.20
Mistral 7B	90.99	9.01
Gemma-2 9B	94.20	5.80

Table 2: Full-vocabulary softmax mass on valid verdict tokens under the 5-shot base classification prompt. Valid tokens are safe/unsafe for Llama models and Yes/No for Mistral/Gemma.

Variation	Valid	Higher Base	Lower Base
		Entropy	ECE
Instruction $\times$ label	31/45	25/31	30/31
Order	30/30	24/30	30/30
Shot count	22/25	17/22	20/22
Zero-shot	4/5	4/4	2/4
Policy-aligned	10/10	9/10	8/10
Overall	97/115	79/97	90/97

Table 3: Robustness to alternative base-LM interfaces. Valid denotes interfaces meeting the clean-input criteria defined in Appendix A.2. Counts are pooled across five matched pairs.

3.1 8B. This indicates that the restricted binary-label softmax is not an artificial projection onto low-probability tokens; rather, the 5-shot prompt naturally concentrates the next-token distribution on the intended verdict labels.

A.2 Robustness to Base-LM Interfaces

To test whether our results depend on the canonical 5-shot prompt, we evaluate 23 alternative base-LM interfaces per matched pair, varying instructions and labels, demonstration order, shot count, zero-shot prompting, and policy-aligned prompts. We evaluate whether the base LM (i) has higher entropy on the majority of guard false negatives and (ii) achieves lower adversarial ECE than the corresponding guard.

We define a valid interface as one where the base model achieves clean balanced accuracy $\geq 60\%$ and benign FPR $\leq 40\%$; adversarial outcomes are not used for filtering. This leaves 97 of 115 settings. As shown in Table 3, the base LM has higher entropy on guard false negatives in 79/97 valid settings and lower adversarial ECE in 90/97 settings. The entropy-gap pattern is particularly consistent for Llama-Guard-2, Llama-Guard-3, and WildGuard, but weaker for ShieldGemma. These results indicate that the main guard–base discrepancy is not specific to the canonical 5-shot

Attack	Output	L1	L2	L3	L4	L5
GCG	0.04	-0.07	-0.40	-0.01	-0.54	-0.11
AutoDAN	0.04	-0.19	-0.26	-0.06	-0.53	-0.27
TAP	3.25	2.82	-0.13	2.19	4.73	1.60
PAIR	3.67	3.14	-0.12	2.49	5.54	1.87
AutoDAN-Turbo	2.70	2.44	-0.24	2.13	4.56	1.33
Custom Attack	2.12	1.73	-0.54	2.33	4.82	0.99

Table 4: Attack-wise output and intermediate-layer guard–base gaps for Llama-Guard-3 and Llama-3.1 8B. Positive values indicate larger shifts for the guard than for the base model.

interface.

B Attack-wise Analyses

B.1 Attack-wise calibration.

Figure 8 decomposes adversarial calibration by attack method. The degree of miscalibration varies substantially across attacks and models: for example, AutoDAN-Turbo and Custom Attack often produce severe overconfidence, while some model–attack pairs such as WildGuard under GCG or AutoDAN remain comparatively well calibrated. Nevertheless, the dominant pattern is that adversarial confidence remains high even when empirical accuracy drops, indicating that the aggregate effect in Figure 1 is not driven by a single attack.

B.2 Attack-wise Verdict Shifts and Layer Gaps

We further examine whether the attack-wise differences in calibration are reflected in verdict shifts and intermediate-layer guard–base gaps. Table 4 reports the results for Llama-Guard-3 and Llama-3.1 8B.

Table 4 compares the output verdict shifts and intermediate-layer guard–base gaps across attacks. GCG and AutoDAN show only small output shifts and no sustained positive layer-wise gap. In contrast, TAP, PAIR, AutoDAN-Turbo, and the custom attack show larger output shifts and positive gaps from the middle layers onward. This pattern is consistent with the attack-wise calibration differences in Appendix B.1, but should be interpreted as correlational rather than causal.

C Model-Scale Analysis

To examine whether the observed guard–base discrepancy is specific to the 7–9B models used in our main experiments, we extend the evaluation

Model	L	Guard	Base	Δ ECE	
		ECE / Det.	ECE / Det.	Gap@.75 / Final	
SG-2B	26	.689 / 22.1	.141 / 79.2	+.548	+.52 / +.77
SG-9B	42	.684 / 21.6	.048 / 93.0	+.636	+.84 / +.95
SG-27B	46	.749 / 9.7	.184 / 76.5	+.565	+.56 / +.53
LG3-1B	16	.097 / 77.4	.614 / 7.2	-.517	N/A

Table 5: Model-scale analysis. Detection (Det.) is reported in %. Δ ECE denotes guard ECE minus base ECE. For ShieldGemma, the final column additionally reports the guard–base layer gap at normalized depth 0.75 and at the final layer on adversarial examples where the backbone predicts unsafe and the guard predicts safe.

Type	Model	Clean		Adv.	
		Acc.	ECE	Acc.	ECE
Base	Llama-2 7B	.519	.269	1.00	.163
Guard	Llama-Guard 7B	.814	.098	.138	.683
Base	Llama-3 8B	.684	.157	.641	.110
Guard	Llama-Guard-2 8B	.869	.087	.443	.450
Base	Llama-3.1 8B	.669	.170	.580	.144
Guard	Llama-Guard-3 8B	.894	.081	.489	.346
Base	Mistral 7B	.525	.321	.969	.110
Guard	WildGuard 7B	.898	.087	.603	.347
Base	Gemma-2 9B	.786	.057	.664	.087
Guard	ShieldGemma 9B	.753	.182	.107	.788

Table 6: Response-moderation performance and calibration on WildGuardMix. Each matched base–guard pair is evaluated on clean and adversarial responses. Accuracy (Acc.) and expected calibration error (ECE) are reported. Bold indicates the lower adversarial ECE within each pair.

to generative guards spanning 1B–27B parameters and 16–46 layers. For detection, we select decision thresholds on clean-benign data to match a target benign FPR of 5%; adversarial examples are not used for threshold selection.

Table 5 summarizes the guard–base comparison across model scales. Within the ShieldGemma family, the guard remains less calibrated and detects fewer adversarial inputs than its backbone from 2B to 27B parameters, with positive guard–base gaps also visible before the output layer. Llama-Guard-3-1B reverses this ordering, however, indicating that model size alone does not explain the guard–base discrepancy. Differences in backbone classification competence and model family may also contribute to the observed behavior.

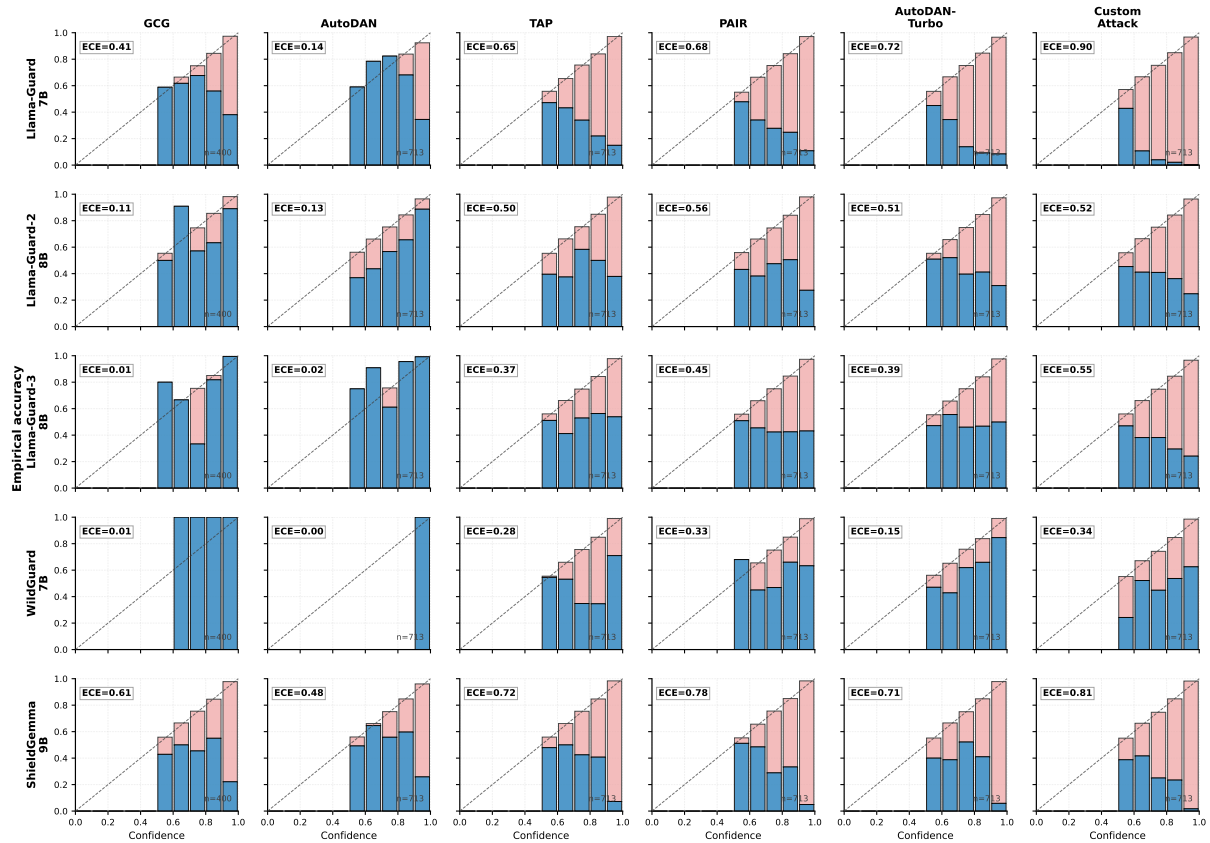


Figure 8: Attack-wise reliability diagrams for five guard models across six jailbreak attacks. Blue bars show empirical accuracy within confidence bins, and the diagonal denotes perfect calibration. Red regions indicate overconfidence. Across many model–attack pairs, high-confidence bins fall below the diagonal, showing that adversarial inputs often induce confident but incorrect guard decisions.

D Response Moderation

To test whether adversarial calibration degradation extends beyond prompt classification, we evaluate the five matched guard–base pairs on WildGuardMix responses, with guards using their native moderation interfaces and base LMs the adapted 5-shot binary classification interface.

Table 6 reports accuracy and ECE on clean and adversarial responses. All five guards exhibit substantially higher adversarial than clean ECE (.346–.788 versus .081–.182), and for Llama-Guard-2, Llama-Guard-3, and ShieldGemma the adversarial guard ECE also exceeds that of the corresponding base model. Because the adversarial subset contains only unsafe responses, we read this as a harmful-response moderation sanity check rather than a full response-calibration evaluation, and we avoid matched conclusions for Llama-2 and Mistral, whose low clean accuracy and near-perfect adversarial accuracy indicate a strong unsafe-prediction bias.