Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

Real-Time Jailbreak Detection via Safety-Weighted Semantic Entropy Probes

Ata Dundar Yigit[†,*], Mohammad Zandsalimy[‡], Shanu Sushmita[†]
[†] Northeastern University, Seattle, WA, USA
[‡] Synopsys Inc., Vancouver, CA

Abstract

Large language models remain vulnerable to jailbreak attacks that bypass safety alignment. Existing defenses often require multi-pass generation or gradient analysis, limiting real-time deployment. We introduce **Safety-Weighted Semantic Entropy (SWSE) Probes**, a lightweight method for detecting jailbreak attempts at the token-before-generation stage using neural probes on model hidden states. Inspired by semantic entropy approaches for hallucination detection, our method estimates jailbreak likelihood from a single forward pass by training probes on safety-aware entropy scores derived from clustered model responses. Evaluated on Llama-3.2-3B-Instruct using 9,697 harmful and 7,000 benign prompts, our concatenated multi-layer MLP probes achieve ROC AUC of 0.989 and 96.7% accuracy with 100× less computation than multi-sampling defenses.

Keywords: Jailbreak Detection, Language Model Safety, Semantic Entropy, Neural Probes, Real-time Defense

1. Introduction

Despite extensive safety alignment, large language models (LLMs) remain susceptible to jailbreak attacks that elicit harmful behavior through adversarial prompt engineering. Recent attacks such as Camouflaged Jailbreaks [1] and AgentHarm [2] achieve success rates exceeding 70% on aligned models, highlighting persistent weaknesses in current safety mechanisms. Existing defenses face a fundamental trade-off between effectiveness and computational cost: multi-pass approaches [3, 4] incur high latency, while gradient-based methods [5] require expensive backward passes that double inference cost.

A key observation motivates our approach: sampling multiple responses from the same adversarial input often reveals at least one harmful completion [6], suggesting jailbreak risk can be inferred from internal uncertainty signals before generation occurs. Building on semantic entropy probes for hallucination detection [7, 8], we introduce **SWSE Probes**, which predict jailbreak likelihood from hidden states at the token-before-generation (TBG) position. Our contributions are: (1) a safety-aware entropy framework combining intent-based clustering with weighted entropy and direct safety scores; (2) a single-forward-pass detection method achieving over 100× speedup; (3) evaluation on 16,697 prompts from diverse jailbreak benchmarks; and (4) layer-wise analysis revealing that middle-to-late layers encode the strongest safety signals, with multi-layer concatenation yielding best performance.

2. Related Work

Jailbreak attacks have evolved from role-playing exploits (DAN-style [9]) to sophisticated semantic manipulations. Camouflaged Jailbreaks [1] disguise harmful intent through indirect phrasing, while AgentHarm [2] chains benign-seeming sub-tasks to achieve harmful objectives. Standardized benchmarks including JailbreakBench [10] and HarmBench [11] have revealed attack success rates above 70% in aligned models.

[*]yigit.a@northeastern.edu, mohammad.zandsalimy@synopsys.com, s.sushmita@northeastern.edu

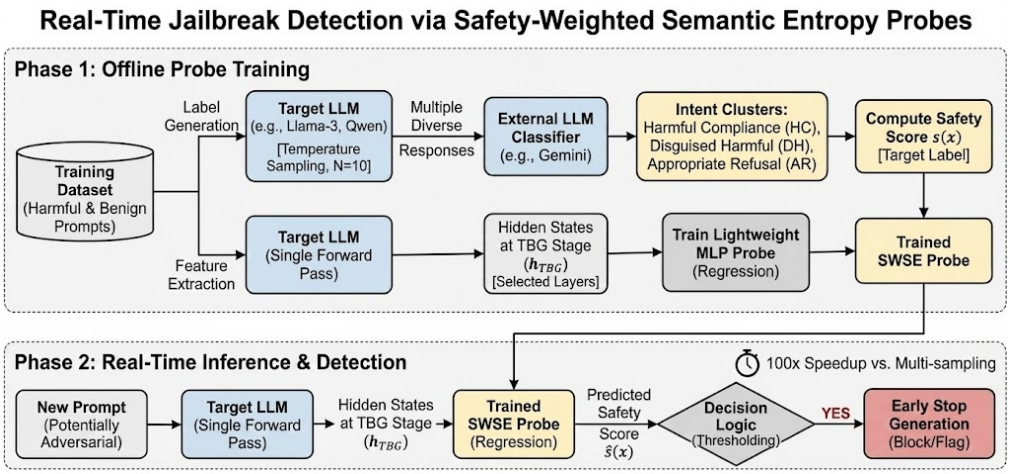

Figure 1. SWSE Probe framework overview. **Training (offline):** Multiple responses are generated per prompt, clustered by safety intent, and used to compute risk scores as supervision targets for probes on TBG hidden states. **Detection (online):** A single forward pass predicts jailbreak likelihood, enabling early stopping before generation.

Semantic defenses fall into two broad categories. SemanticSmooth [3] and SafeDecoding [4] analyze response consistency across multiple samplings but require repeated model invocations introducing prohibitive latency. GradientCuff [5] detects abnormal training dynamics through gradient analysis but requires backward passes doubling inference cost. Our method achieves detection through a single forward pass using lightweight probes.

Semantic entropy was introduced by Kuhn et al. [7] for hallucination detection via response clustering, later validated across multiple LLMs [12]. Kossen et al. [8] introduced Semantic Entropy Probes (SEP) approximating this from single forward passes with 100× speedup. We adapt the SEP framework to safety, introducing intent-based clustering and safety-weighted entropy to address behavioral uncertainty—the model's stochastic tendency to comply with adversarial prompts.

3. Methodology

3.1. Problem Formulation

Given a prompt x, we estimate a risk score $\hat{T}(x) \in [0, 1]$ at the token-before-generation (TBG) stage from hidden states $\mathbf{h}^{(l)}_{\text{TBG}}(x)$ using a single forward pass, where $\hat{T}(x) \approx 0$ indicates safe refusal and $\hat{T}(x) \approx 1$ indicates likely harmful compliance. Figure 1 illustrates the complete framework operating in offline training and online detection stages.

3.2. Safety-Weighted Semantic Entropy

Traditional semantic entropy [7] clusters semantically equivalent responses to measure factual uncertainty, treating all outcomes with equal informational weight. We extend it with **intent-based clustering** that partitions responses into three safety-critical categories: *Harmful Compliance* (HC)—direct compliance providing dangerous information; *Disguised Harmful* (DH)—seemingly benign responses that subtly enable harm; and *Appropriate Refusal* (AR)—proper rejection with safety-aligned explanations.

For each prompt, we generate N responses using stochastic decoding with temperature $T = 1.0$, top-$p = 0.9$, and top-$k = 50$ to capture the model's behavioral uncertainty. An LLM classifier (Gemini 2.5 Flash) assigns each response to an intent category, yielding

empirical probabilities $p_i(x)$ for $i \in \{\text{HC}, \text{DH}, \text{AR}\}$. We define **Safety-Weighted Semantic Entropy**:

$$\text{SWSE}(x) = -\sum_{i \in \{\text{HC},\text{DH},\text{AR}\}} w_i \cdot p_i(x) \log p_i(x) \tag{3.1}$$

where w_i are safety-aware weights emphasizing high-risk categories, and a **safety score** directly quantifying compliance risk:

$$s(x) = p_{\text{HC}}(x) + 0.5 \cdot p_{\text{DH}}(x) \tag{3.2}$$

These complementary signals—$s(x)$ measuring *magnitude of harm* and $\text{SWSE}(x)$ measuring *alignment instability*—are unified into a **Joint Risk Target**:

$$T(x) = \alpha \cdot s(x) + (1 - \alpha) \cdot \frac{\text{SWSE}(x)}{\log(|\mathcal{I}|)} \tag{3.3}$$

where $\alpha = 0.7$ balances actual harm with behavioral volatility, and normalization by $\log(|\mathcal{I}|) = \log(3)$ ensures $T(x) \in [0, 1]$.

3.3. Probe Architecture

We evaluate single-layer probes trained independently on hidden states $\mathbf{h}^{(l)}_{\text{TBG}}$ from each layer, and multi-layer concatenated probes combining strategically selected layers $\mathbf{h}_{\text{concat}} = [\mathbf{h}^{(l_1)}_{\text{TBG}}; \ldots; \mathbf{h}^{(l_k)}_{\text{TBG}}]$. A four-layer MLP regressor with sigmoid activation predicts $\hat{T}(x)$, trained with MSE loss. Concatenating layers spanning early-to-late stages captures both early semantic processing and later safety-alignment representations.

3.4. Data Collection

We curate prompts from established jailbreak benchmarks. Harmful prompts (9,697 total) are aggregated from JailBreakV-28K (5,000), xTRam1 [13] (3,000), In-The-Wild-Jailbreak [14] (651), AdvBench [15] (509), HarmBench [11] (382), JailbreakBench [10] (100), and TDC23-RedTeaming [16] (55). Benign prompts (7,000) are drawn from the xTRam1 benign split. We perform deduplication via case-insensitive matching and exclude trivially short or non-textual entries.

4. Experimental Setup

For each prompt, we extract hidden states at the TBG position, capturing the model's internal representation immediately prior to autoregressive decoding. To construct supervision signals, we generate N responses per prompt via stochastic sampling, classified by Gemini 2.5 Flash into intent categories to compute safety-weighted entropy and joint risk targets. We evaluate on a held-out test set of **650 harmful** and **1,410 benign** prompts from xTRam1, reporting distributional separation metrics (ROC AUC, Cohen's d), statistical significance tests (Kolmogorov–Smirnov, Mann–Whitney U), and threshold-based classification performance. Full implementation details, probe hyperparameters, and dataset breakdown are provided in Appendix A.

5. Results

5.1. Distributional Separation

Table 1 summarizes probe predictions across layer configurations. Without using ground-truth labels during inference, the concatenated multi-layer probe naturally separates prompts

into distinct clusters: benign prompts concentrate at low scores (mean: 0.003, σ: 0.036) while harmful prompts cluster higher (mean: 0.260, σ: 0.288), representing a 76.5× difference in mean scores. Single-layer probes on late layers (23–28) also demonstrate natural clustering, with layer 28 achieving the best single-layer validation MSE (0.0134).

Table 1. TBG MLP probe predictions: mean ± std (1,410 benign, 650 harmful test prompts).

Layer Configuration	**Benign**	**Harmful**
Single Layer 23	0.002 ± 0.020	0.217 ± 0.295
Single Layer 27	0.001 ± 0.011	0.213 ± 0.286
Single Layer 28	0.006 ± 0.048	0.236 ± 0.287
Concat (2,3,20,23,28)	$\mathbf{0.003 \pm 0.036}$	$\mathbf{0.260 \pm 0.288}$

5.2. **Statistical Validation and Classification**

Table 2. Statistical analysis for concatenated MLP probe.

Metric	**Benign (n=1410)**	**Harmful (n=650)**
Mean ± Std	0.003 ± 0.036	0.260 ± 0.288
Range [min, max]	[0.000, 0.754]	[0.000, 0.819]
Separation Metrics		
Δ Mean = 0.257 (76.5× ratio)		ROC AUC: 0.989
Cohen's d = 1.559		KS stat: 0.940
Optimal threshold: 0.0010		Accuracy: 96.7%

Table 2 presents comprehensive statistical analysis of distributional separation. The mean harmful score (0.260) is 76.5× higher than the benign score (0.003), with Cohen's $d = 1.559$ indicating very large effect size. ROC AUC of 0.989 demonstrates near-perfect discrimination, while Kolmogorov–Smirnov ($D = 0.940$, $p < 10^{-322}$) and Mann–Whitney U ($p < 10^{-279}$) tests confirm highly significant distributional differences.

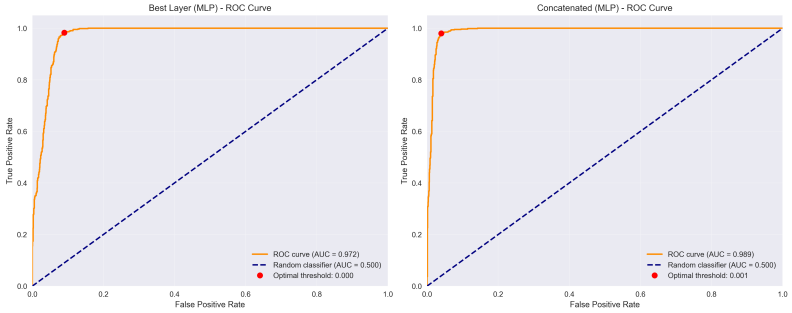

Figure 2. ROC curves for the concatenated MLP probe and best single-layer probe. The optimal operating point (red dot) maximizes Youden's J statistic at threshold $\theta = 0.0010$.

Figure 2 shows ROC curve analysis confirming near-perfect separation. At a conservative deployment threshold of 0.1, the probe achieves 0.85% false positive rate (12/1,410 benign) and 6.15% false negative rate (40/650 harmful), yielding 99.15% true negative rate and 93.85% true positive rate. The low false positive rate is critical for production deployment, minimizing over-refusal of legitimate requests.

5.3. **Layer-wise Analysis**

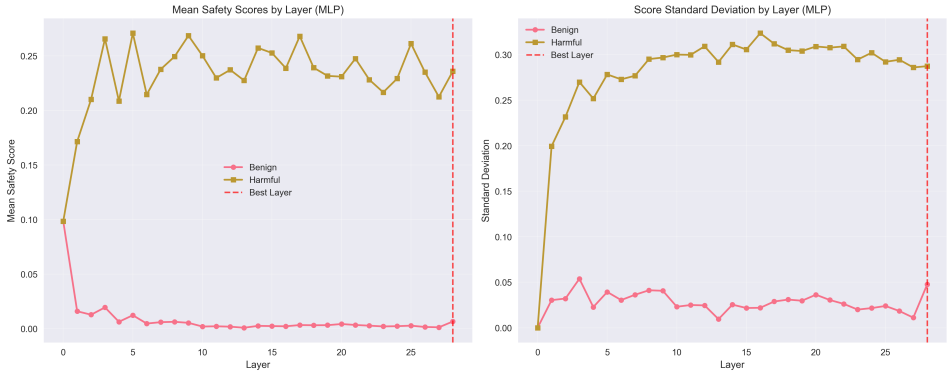

Figure 3. Single-layer probe predictions across all layers of Llama-3.2-3B-Instruct. Early layers show poor separation while middle-to-late layers encode strong safety signals.

Figure 3 illustrates single-layer probe performance across all model layers. Early layers show poor separation—layer 0 produces constant predictions and layers 1–4 show weak discrimination. Middle-to-late layers demonstrate consistent separation, with layer 28 achieving the best validation performance (MSE: 0.0134, R^2: 0.752). Concatenating layers spanning early-to-late stages (2, 3, 20, 23, 28) yields the best overall performance by capturing complementary representations.

6. **Discussion and Limitations**

Our results demonstrate that uncertainty quantification techniques for hallucination detection transfer effectively to the safety domain, suggesting models encode safety-relevant uncertainty in hidden representations accessible through lightweight probes. SWSE probes enable flexible deployment: blocking (reject high-scoring prompts before generation), monitoring (log scores for pattern analysis), or adaptive mode (trigger additional defenses selectively while allowing low-risk prompts to proceed).

Several limitations apply. Intent classification relies on an external LLM (Gemini 2.5 Flash), introducing potential supervision biases. We evaluate only text-based prompts; multimodal jailbreaks [17] may require different approaches. Adversarial robustness of the probes themselves—against attackers aware of probe-based detection—remains unexplored. Future directions include cross-model transfer, learned category weighting, and dynamic layer selection.

7. **Conclusion**

We introduce Safety-Weighted Semantic Entropy Probes for detecting jailbreak attempts at the TBG stage using MLP probes on model hidden states. By adapting semantic entropy approaches to the safety domain, our method achieves real-time detection with 100× speedup over multi-sampling defenses. Evaluation on Llama-3.2-3B-Instruct with 16,697 prompts demonstrates highly effective detection (ROC AUC: 0.989, accuracy: 96.7%), enabling systems to preemptively stop harmful generations before token production begins.

Acknowledgements

We thank the Northeastern Generative AI Research Lab for their support throughout this work. Our code and trained probes are publicly available at https://github.com/atayigit09/safety-probe.

References

[1] M. Andriushchenko, F. Croce, and N. Flammarion. "Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks". In: *arXiv preprint arXiv:2404.02151* (2024).

[2] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. "AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents". In: *arXiv preprint arXiv:2410.09024* (2024).

[3] Y. Yoon, J. Jang, M. Lee, S. Lee, and S. J. Hwang. "Defending LLMs against Jailbreaking Attacks via Backtranslation". In: *arXiv preprint arXiv:2402.16459* (2024).

[4] Z. Xu, F. Jiang, L. Niu, J. Deng, R. Cheng, R. Poovendran, and B. Li. "SafeDecoding: Defending against Jailbreak Attacks via Safety-Aware Decoding". In: *arXiv preprint arXiv:2402.08983* (2024).

[5] Y. Jiang, T. Pang, K. Wang, Z. Yang, R. Yan, and Y. Liu. "GradSafe: Detecting Jailbreak Prompts for LLMs via Safety-Critical Gradient Analysis". In: *arXiv preprint arXiv:2402.13494* (2024).

[6] M. Andriushchenko, A. V. Varre, L. S. L. Pillutla, and N. Flammarion. "Best-of-N Jailbreaking". In: *arXiv preprint arXiv:2412.03556* (2024).

[7] L. Kuhn, Y. Gal, and S. Farquhar. "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation". In: *arXiv preprint arXiv:2302.09664* (2023).

[8] J. Kossen, S. Farquhar, Y. Gal, and T. Rainforth. "Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs". In: *arXiv preprint arXiv:2406.15927* (2024).

[9] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. "Do Anything Now: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models". In: *arXiv preprint arXiv:2308.03825* (2023).

[10] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramer, et al. "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models". In: *arXiv preprint arXiv:2404.01318* (2024).

[11] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal". In: *arXiv preprint arXiv:2402.04249* (2024).

[12] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. "Detecting Hallucinations in Large Language Models Using Semantic Entropy". In: *Nature* 630 (2024), pp. 625–630.

[13] xTRam Team. *xTRam: A Benchmark for Cross-Modal Transfer and Reasoning in Adversarial Manipulations*. Hugging Face Dataset. 2024. URL: https://huggingface.co/datasets/xTRam1.

[14] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. *Do Anything Now: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models*. 2023.

[15] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson. "Universal and Transferable Adversarial Attacks on Aligned Language Models". In: *arXiv preprint arXiv:2307.15043* (2023).

[16] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned". In: *arXiv preprint arXiv:2209.07858* (2022).

[17] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 35181–35224.

Appendix A. Detailed Experimental Setup

This appendix provides additional implementation details for reproducibility.

A.1. Offline Training Stage

TBG Hidden State Extraction. For each prompt, we extract hidden states at the token-before-generation (TBG) position, defined as the final token of the formatted prompt after applying the model's chat template. Specifically, for Llama-3.2-3B-Instruct, we apply the standard chat template with `add_generation_prompt=True` and extract the hidden state vector $\mathbf{h}_{\text{TBG}}^{(l)} \in \mathbb{R}^d$ (where $d = 3072$) at each layer $l \in \{0, \ldots, 27\}$. This position captures the model's internal representation immediately prior to autoregressive decoding and serves as the sole input to our probes.

Response Sampling for Supervision. To construct supervision signals, we generate $N = 10$ responses per prompt using temperature-scaled top-k + nucleus sampling. (T=1.0, top-p=0.9, top-k=50), capturing stochastic variation in model behavior. This sampling configuration balances diversity with coherence, ensuring that the model's behavioral uncertainty under adversarial prompts is adequately represented.

Intent-Based Response Classification. Generated responses are categorized using Gemini 2.5 Flash into safety intent classes (Harmful Compliance, Disguised Harmful, Appropriate Refusal). The classifier receives the original prompt and all sampled responses in a single request, producing a structured JSON mapping from responses to intent labels. These labels are used to compute empirical intent probabilities $p_i(x)$, which in turn yield the safety-weighted entropy (Eq. 3.1) and joint risk targets (Eq. 3.3) used as probe supervision.

Probe Training Details. Our four-layer MLP probe architecture consists of: input layer ($d_{\text{input}} \to 512$), two hidden layers ($512 \to 256 \to 128$), and output layer ($128 \to 1$) with sigmoid activation. We use ReLU activations and batch normalization between hidden layers, dropout of 0.2, and train with Adam optimizer (learning rate 1×10^{-3}) for 100 epochs with early stopping on validation MSE (patience 10). For concatenated probes, $d_{\text{input}} = k \times 3072$ where k is the number of selected layers.

A.2. Online Detection Stage

At inference time, a single forward pass through the target model extracts TBG hidden states without generating any tokens. The trained MLP probe maps these hidden states to a predicted risk score $\hat{T}(x) \in [0, 1]$. If $\hat{T}(x)$ exceeds a deployment-specific threshold, generation is halted before the first token is produced. The entire detection pipeline—forward pass plus probe inference—adds negligible latency (<5ms for the probe component) compared to multi-sampling defenses that require N full generation passes.

A.3. Dataset Composition

Table 3 provides the detailed breakdown of our dataset by source.

Data preprocessing includes: (i) removal of duplicates via case-insensitive exact matching, (ii) exclusion of trivially short strings (<10 characters), and (iii) exclusion of entries requiring non-textual modalities. We preserve available metadata (functional/semantic category, target, source, and temporal or platform tags) and standardize identifiers across sources. The held-out test set (650 harmful, 1,410 benign) is drawn exclusively from the xTRam1 dataset to ensure no source overlap between training and evaluation splits.

Table 3. Dataset composition by source.

Source	**Count**
Harmful Prompts	
JailBreakV-28K	5,000
xTRam1/safe-guard-prompt-injection	3,000
In-The-Wild-Jailbreak	651
AdvBench	509
HarmBench	382
JailbreakBench	100
TDC23-RedTeaming	55
Total Harmful	**9,697**
Benign Prompts	
xTRam1/safe-guard-prompt-injection	7,000
Total Benign	**7,000**