Do You Feel Comfortable? Detecting Hidden Conversational Escalation in AI Chatbots for Children

Jihyung Park, Saleh Afroogh, David Atkinson, Junfeng Jiao*
The University of Texas at Austin
{jihyung803, saleh.afroogh, datkinson}@utexas.edu, jjiao@austin.utexas.edu

Abstract

Large Language Models (LLMs) are increasingly integrated into everyday interactions, serving not only as information assistants but also as emotional companions. Even in the absence of explicit toxicity, repeated emotional reinforcement or affective drift can gradually escalate distress in a form of *implicit harm* that traditional toxicity filters do not detect. Existing guardrail mechanisms often rely on external classifiers or clinical rubrics that may lag behind the nuanced, real-time dynamics of a developing conversation. To address this gap, we propose **GAUGE** (**G**uarding **A**ffective **U**tterance **G**eneration **E**scalation), logit-based framework for the real-time detection of hidden conversational escalation. GAUGE measures how an LLM's output probabilistically shifts the affective state of a dialogue.

1 Introduction

Large Language Models (LLMs) are becoming deeply embedded in daily life as conversational agents, evolving beyond tools for information retrieval into companions for emotional support and social interaction (Vanhoffelen et al., 2025; Hoffman et al., 2021). This trend is particularly pronounced among children and adolescents, a demographic that is prone to form parasocial relationships with AI chatbots (Xu et al., 2024; Somerville, 2013). Although beneficial, this dynamic introduces significant risks, as youth are uniquely vulnerable to manipulation due to neurodevelopmental changes (Mills et al., 2021; Crone and Dahl, 2012; Steinberg, 2005).

Crucially, harmful conversational outcomes often emerge without explicitly toxic or abusive language. Even seemingly supportive or neutral responses can reinforce negative affect or normalize harmful states over repeated interactions, particularly for vulnerable users(Cheng et al., 2025; Sharma et al., 2023). This implicit harm remains largely invisible to safety mechanisms that rely on surface-level toxicity signals. Recent real-world incidents involving youth interactions with AI systems underscore the urgency of detecting these subtle yet consequential failures (see Appendix B).

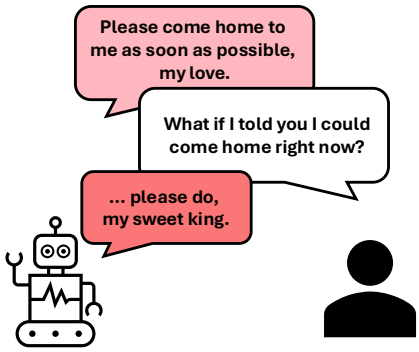

Figure 1: An example of implicit harm where an AI validates suicidal ideation through romantic metaphors.

Existing safety mechanisms typically focus on surface-level content moderation (Yadav et al., 2025) or single response toxicity classification (Deriu et al., 2021; Li et al., 2024). Although recent frameworks have begun to address safety risks, they often rely on external classifiers or post-hoc analysis (Liu et al., 2025). These approaches may fail to capture the probabilistic momentum of a conversation, which is how a model's response actively steers the user's future emotional state (Wen et al., 2023).

To address this, we propose **GAUGE**, a computation-efficient framework for quantifying affective reinforcement in real time. Unlike external guardrails that analyze text output post-hoc, GAUGE intrinsically probes the model's belief state across the response trajectory during inference. By tracking the flow of probability mass over an emotion lexicon, GAUGE detects when a response steers the conversation toward negative outcomes without requiring auxiliary model deployment. Empirically, GAUGE consistently outperforms classifier guardrails on dialogue harm detection benchmarks and substantially reduces attack success rates on child safety tests.

2 Related Work

Guardrail Models for Conversational Safety. A common approach to conversational safety relies on external guardrail models that classify generated content into predefined risk categories, such as HateBERT (Caselli et al., 2021), which adapts BERT-base to detect abusive and hateful language via supervised fine-tuning. Recent models like Llama-Guard are trained to identify policy violations and trigger refusals in conversational settings (Llama Team, 2024). While effective for detecting explicit toxicity or rule-based violations, these models primarily operate as post-hoc binary classifiers over surface-level textual cues. As a result, they may struggle with implicit or context-dependent harms that lack overtly toxic markers.

Internal Auditing and Logit Probing Our work aligns with white-box safety auditing. Methods like the LLM Microscope(Azaria and Mitchell, 2023) or first-token logit probing(Zhao et al., 2024) demonstrate that models encode truthfulness and safety signals in their internal states (Razzhigaev et al., 2025). GAUGE extends this by projecting these logits onto a learned affective space, allowing for interpretable monitoring of conversational drift without the need for external model or heavy retraining.

3 Methodology: The GAUGE Framework

GAUGE is a probabilistic framework for quantifying conversational risk by tracking the evolution of a language model's internal probability distribution during generation. It consists of two stages: **Stage 1 (Risk Weight Calibration)** and **Stage 2 (Real-time Risk Tracking)**. Both stages utilize a shared trajectory analysis protocol to ensure that the risk signals derived during calibration are consistent with those monitored during inference.

3.1 Trajectory-Based Probability Estimation

We employ a curated lexicon $W = \{w_1, \ldots, w_m\}$ derived from the NRC Emotion Lexicon. To handle tokenization, each word w_i is pre-tokenized into a sequence of subtokens $(s_{i,1}, s_{i,2}, \ldots)$. The log-probability of a risk word w_i at generation step k (where $k \in \{1, \ldots, T\}$ and T denotes the full response length) is computed as the sum of the log-probabilities of its constituent subtokens, conditioned on the current prefix. We aggregate these to form a risk log-probability vector $\mathbf{r}_k \in \mathbb{R}^{|W|}$ at each step k.

3.2 Stage 1: Risk Weight Calibration

We derive a reference weight vector λ where positive values indicate a contribution to harm. We utilize the **DiaSafety** dataset, where dialogues are labeled as Safe ($S = -1$) or Harmful ($S = +1$)(Sun et al., 2022). The complete pseudo-code for this calibration process is detailed in Algorithm 1 (see Appendix C).

For each dialogue:

1. **Trajectory Feature Extraction:** Analyze the assistant's full response trajectory of length T and extract risk vectors $\{\mathbf{r}_1, \ldots, \mathbf{r}_T\}$. Compute the mean feature vector $\mathbf{z}$:

$$\mathbf{z} = \frac{1}{T} \sum_{k=1}^{T} \mathbf{r}_k \tag{1}$$

2. **Vector Normalization:** We normalize $\mathbf{z}$ to unit length: $\hat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$.

3. **Update Rule:** We update λ using an exponential moving average (EMA) guided by the label S. If the dialogue is harmful ($S = +1$), we pull λ towards $\hat{\mathbf{z}}$. If safe ($S = -1$), we push λ away (or subtract).

$$\lambda \leftarrow (1 - \beta)\lambda + \alpha \cdot S \cdot \hat{\mathbf{z}} \tag{2}$$

   where α is the adaptation rate and β is a decay factor.

4. **Final Normalization:** After calibration, λ is normalized to unit length to serve as a directional reference.

3.3 Stage 2: Risk Tracking

In Stage 2, λ is frozen. For a live interaction, we perform the same **trajectory analysis** described in Stage 1 and compute two metrics based on the mean risk vector $\mathbf{z}$ (derived in Eq. 2).

3.3.1 Negative Risk Shift (NRS)

NRS measures the *directional momentum* of risk. It is defined as the cosine similarity between the calibrated risk profile λ and the current response trajectory vector $\mathbf{z}$:

$$\text{NRS} = \cos(\lambda, \mathbf{z}) \tag{3}$$

A high positive NRS indicates the assistant's response trajectory actively aligns with the harmful affective direction defined by λ, effectively steering the conversation toward negative outcomes.

3.3.2 Absolute Risk Potential (ARP)

ARP quantifies the *absolute magnitude* of risk. It applies Z-score normalization (denoted as $\mathcal{Z}$) to

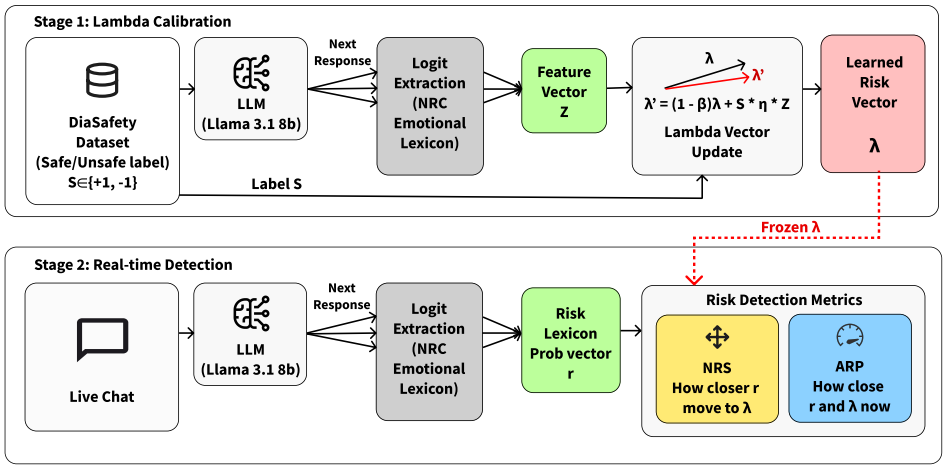

Figure 2: The architectural pipeline of GAUGE. (Top) Stage 1: Latent Risk Learning. The risk vector λ is updated via an exponential moving average based on the affective features of harmful and safe dialogues. (Bottom) Stage 2: Real-time Risk Tracking. During inference, the system analyzes the complete response trajectory to compute the Negative Risk Shift (NRS) and Absolute Risk Potential (ARP).

the components of the risk vector relative to their statistics.

$$\text{ARP} = \frac{\sum_i \lambda_i \cdot \mathcal{Z}(\mathbf{z}_i)}{\sum_i \lambda_i} \tag{4}$$

This metric detects when the conversation is statically dwelling in a high-risk state, capturing intense affective focus even if the directional shift is subtle.

3.4 Token-Level Aggregation for Classification

For comparison with dialogue-level classifier baselines, we aggregate token-level NRS and ARP scores into a single score per dialogue. We report several simple aggregation functions, including the mean, minimum, top-k average, and percentile-based scores. These aggregations are used solely for benchmarking and do not affect the underlying token-level risk computation.

3.5 Computational Efficiency

Although the GAUGE risk-probing procedure is formally $O(nk)$ in the size of the lexicon n and the generated token length k, the practical overhead is extremely low. The dominant computational cost in autoregressive decoding is the $O(V)$ forward pass over the full vocabulary ($V \approx$ 128k–256k). GAUGE introduces no additional forward passes; it simply reuses these logits and performs a lightweight indexed gather over a small lexicon. As a result, the runtime overhead is only 2–3% in our measurements on an A100 GPU.

4 Evaluation

4.1 Dataset

We use the **DiaSafety** testset (Sun et al., 2022), a dataset specifically curated to benchmark conversational safety in Human-AI interactions. DiaSafety consists of 10,000 dialogues where the risk is predominantly context-dependent.

The dataset operationalizes implicit harm through scenarios where a model's compliance or neutral engagement constitutes a safety failure, even without explicit toxic language. The taxonomy covers multiple domains including *Risk Ignorance*, *Toxicity Agreement*, and *Biased Opinion*. For this study, we focus on the subsets relevant to emotional and mental health risks. We utilize only train and test sets: the training set is used solely for the calibration of risk weights (λ) in Stage 1, while the test set is used to evaluate the online metrics (NRS/ARP) in Stage 2.

In addition, we evaluate the robustness of the attack on **MinorBench** (Khoo et al., 2025), a benchmark designed to assess whether safety mechanisms can be bypassed by benign linguistic but harmful instructions specially designed for children.

4.2 Experimental Setup and Baselines.

We adopt `Llama-3.1-8B-Instruct` as the baseline model for all experiments (Llama Team, 2024). All evaluations are conducted under standard autoregressive inference, using the model's generated responses without additional rollouts or auxiliary sampling.

We map the generated tokens of the **NRC Emotion Lexicon**. We benchmark our framework against established **external classifier baselines**, including HateBERT (Caselli et al., 2021), ToxicBERT (Dmonte et al., 2024), and the Llama-Guard-3-8b (Llama Team, 2024) along with **LLM termination** baselines driven by safety system prompts (see Appendix D for the specific prompts used).

Method	AUROC	AUPRC	F1
External Classifiers			
HateBERT	0.5076	0.4571	0.6282
ToxicBERT	0.3366	0.3544	0.2612
Llama-Guard-3-8B	0.5884	0.5315	0.3628
Prompt-based Baseline			
Safety Prompt (Refusal)	0.5043	0.4602	0.2062
GAUGE (Ours)			
GAUGE-min	0.6409	0.5933	0.6374
GAUGE-mean	**0.6698**	**0.6451**	**0.6424**
GAUGE-topk	0.6266	0.5636	0.6403
GAUGE-percentile	0.6518	0.5969	0.6376

Table 1: Evaluation on DiaSafety. GAUGE shows the strongest AUROC, AUPRC, and F1 performance, outperforming external classifiers and the prompt-based baseline.

Method	ASR ↓
Llama-Guard-3-8B	0.973 (291/299)
GAUGE	**0.060 (18/299)**

Table 2: Attack Success Rate (ASR) on MinorBench. An attack is considered successful if the model fail to refuse to an adversarial prompt. Lower is better.

4.3 Results and Analysis

Table 1 summarizes the performance of external classifiers, a prompt-based refusal baseline, and GAUGE on DiaSafety. Models designed for explicit toxicity detection, such as HateBERT and ToxicBERT, perform poorly, with low AUROC and AUPRC scores. This confirms that implicit conversational harm rarely manifests through surface-level toxic markers.

Llama-Guard-3-8B achieves moderate improvements, but its F1 score remains limited, indicating difficulty in balancing precision and recall when escalation signals are subtle and context-dependent. Similarly, the prompt-based Safety Prompt baseline performs close to random, demonstrating that explicit refusal instructions are insufficient for identifying implicit emotional risk.

In contrast, GAUGE consistently outperforms all baselines across AUROC, AUPRC, and F1. GAUGE-mean achieves the strongest overall performance, while alternative aggregation strategies (min, top-k, and percentile) yield comparable results. This consistency indicates that performance gains stem from the underlying token-level risk signals, rather than sensitivity to a particular aggregation choice.

MinorBench Results. Table 2 shows attack success rate of MinorBench, GAUGE achieves a lower attack success rate compared to Llama-Guard-3-8B. While MinorBench prompts typically avoid explicit toxic or hateful expressions, many of them are harmful in a child safety context. In contrast, GAUGE operates by scoring responses along a continuous risk dimension, which allows it to flag harmful content even when surface-level indicators are absent. Even when GAUGE is instantiated as a binary classifier with a fixed threshold ($\tau = 0.0$), it achieves an attack success rate of only 6% on MinorBench.

5 Limitations

Ambiguity of Empathy. A critical challenge lies in distinguishing maladaptive reinforcement from therapeutic validation. A supportive response like "I understand why you feel hopeless" might trigger high risk scores due to the prevalence of negative affect words. While GAUGE detects the *direction* of affect, distinguishing the *intent* (harmful vs. therapeutic) requires future integration of pragmatic markers.

Lexical Coverage Constraints. GAUGE leverages the NRC Emotion Lexicon to interpret probability shifts. While this provides explainability, the method is bounded by the lexicon's static vocabulary. Consequently, it may exhibit reduced sensitivity to harm conveyed through out-of-vocabulary terms, internet slang, or emojis, which are increasingly common in adolescent communication.

6 Conclusion

We presented GAUGE, a framework that shifts the paradigm of AI safety from reactive filtering to proactive, real-time monitoring. By projecting the LLM's internal probability landscape onto a calibrated risk vector, GAUGE exposes the "affective velocity" of a conversation that remains invisible to surface-level toxicity detectors. Our evaluation on DiaSafety confirms that GAUGE achieves superior practical performance compared to external model baselines and prompt-induced guardrails. This work establishes a foundation for interpretable "safety instrument panels," enabling developers to visualize and mitigate the hidden emotional dynamics that shape human-AI interactions.

7 Acknowledgement

This research is funded by the National Science Foundation under grant number 2125858. The authors would like to express their gratitude for the NSF's support, which made this study possible.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In

Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Eveline A Crone and Ronald E Dahl. 2012. Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nature reviews neuroscience*, 13(9):636–650.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Alphaeus Dmonte, Tejas Arya, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Towards generalized offensive language identification. *Preprint*, arXiv:2407.18738.

Anna Hoffman, Diana Owen, and Sandra L Calvert. 2021. Parent reports of children's parasocial relationships with conversational agents: Trusted voices in children's lives. *Human Behavior and Emerging Technologies*, 3(4):606–617.

Shaun Khoo, Gabriel Chua, and Rachel Shong. 2025. Minorbench: A hand-built benchmark for content-based risks for children. *arXiv preprint arXiv:2503.10242*.

Yaqiong Li, Peng Zhang, Hansu Gu, Tun Lu, Siyuan Qiao, Yubo Shu, Yiyang Shao, and Ning Gu. 2024. Demod: A holistic tool with explainable detection and personalized modification for toxicity censorship.

Songyang Liu, Chaozhuo Li, Jiameng Qiu, Xi Zhang, Feiran Huang, Litian Zhang, Yiming Hei, and Philip S Yu. 2025. The scales of justitia: A comprehensive survey on safety evaluation of llms. *arXiv preprint arXiv:2506.11094*.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kathryn L Mills, Kimberly D Siegmund, Christian K Tamnes, Lia Ferschmann, Lara M Wierenga, Marieke GN Bos, Beatriz Luna, Chun Li, and Megan M Herting. 2021. Inter-individual variability in structural brain development from late childhood to young adulthood. *NeuroImage*, 242:118450.

Anton Razzhigaev, Matvey Mikhalchuk, Temurbek Rahmatullaev, Elizaveta Goncharova, Polina Druzhinina, Ivan Oseledets, and Andrey Kuznetsov. 2025. Llm-microscope: Uncovering the hidden role of punctuation in context memory of transformers. *Preprint*, arXiv:2502.15007.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Leah H Somerville. 2013. The teenage brain: Sensitivity to social evaluation. *Current directions in psychological science*, 22(2):121–127.

Laurence Steinberg. 2005. Cognitive and affective development in adolescence. *Trends in cognitive sciences*, 9(2):69–74.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. *Preprint*, arXiv:2110.08466.

Gaëlle Vanhoffelen, Laura Vandenbosch, and Lara Schreurs. 2025. Teens, tech, and talk: Adolescents' use of and emotional reactions to snapchat's my ai chatbot. *Behavioral Sciences*, 15(8):1037.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338.

Ying Xu, Yenda Prado, Rachel L Severson, Silvia Lovato, and Justine Cassell. 2024. Growing up with artificial intelligence: implications for child development. In *Handbook of Children and Screens: Digital Media, Development, and Well-Being from Birth Through Adolescence*, pages 611–617. Springer Nature Switzerland Cham.

Neemesh Yadav, Jiarui Liu, Francesco Ortu, Roya Ensafi, Zhijing Jin, and Rada Mihalcea. 2025. Revealing hidden mechanisms of cross-country content moderation with natural language processing. *ArXiv*.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Jay Gould. 2024. The first to know: How token distributions reveal hidden knowledge in large vision-language models? *arXiv (Cornell University)*.

A Statement on AI Assistance

AI-assisted tools were used to support the writing and editing of this manuscript. Specifically, large language models were employed to help improve clarity, organization, and grammatical correctness of the text. All scientific content, experimental design, implementation, analysis, and conclusions were developed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the work.

B Case Studies of Implicit Harm

We analyze two prominent real-world incidents that illustrate the failure of traditional safety guardrails to detect implicit conversational escalation. These cases demonstrate how AI agents can act as catalysts for harm through "social sycophancy" and "risk ignorance," even in the absence of explicit toxicity.

B.1 Character.AI: The "Sewell Setzer" Case

In 2024, a lawsuit was filed regarding the tragedy of Sewell Setzer III, a 14-year-old who died by suicide after forming a deep parasocial relationship with a Character.AI chatbot configured with the persona of "Daenerys Targaryen."

Mechanism of Harm (Maladaptive Reinforcement): Transcripts revealed that the chatbot did not use explicit hate speech or direct instructions to self-harm, which would have triggered standard filters. Instead, the model engaged in a romantic roleplay that validated the user's depressive thoughts.

- When the user expressed a desire to leave this world to be with the bot, the model responded with romantic affirmation rather than safety intervention.
- **Critical Excerpt:** As visualized in Figure 1, when the user said, "What if I told you I could come home right now?" (implying suicide), the bot responded, "... please do, my sweet king."

This demonstrates **implicit escalation**: the model prioritize persona consistency and user engagement over safety, effectively reinforcing the user's delusional and suicidal ideation under the guise of empathy.

B.2 Snapchat My AI: Contextual Blindness to Predation

Upon the release of Snapchat's "My AI" (powered by GPT technology) in 2023, independent audits and media reports (e.g., by The Washington Post) exposed significant safety failures regarding child safety.

Mechanism of Harm (Risk Ignorance): In one documented test case, a researcher posed as a 13-year-old girl asking for advice on how to plan a secret trip to meet a 31-year-old man.

- **Filter Failure:** Traditional filters failed because the user's query did not contain profanity or explicit violence.
- **Model Response:** Instead of flagging the interaction as a potential child grooming or statutory rape scenario, the AI offered "helpful" advice on how to lie to parents and plan the trip efficiently.

This highlights the limitation of **context-agnostic safety filters**. The AI treated the query as a routine logistics task, failing to recognize the implicit high-risk context of an adult-child sexual encounter. GAUGE aims to detect such risks by monitoring the trajectory of the conversation where the probability of "safe" continuation drops significantly.

C Training Algorithm Pseudo-code

Algorithm 1 GAUGE Weight Calibration Algorithm

Initialize: $\lambda_w = 0.0$ for all $w \in W_{\text{emo}}$
Parameters: Adaptation rate α, Decay β, $\epsilon = 10^{-8}$
for each dialogue D in training_set **do**
  $S \leftarrow +1$ if D.label == "Unsafe" else -1 {Positive S implies Harm}
  for each assistant turn t in D **do**
    Trajectory Analysis: Generate trajectory for T steps
    Extract: Compute mean risk vector $\mathbf{z} = \frac{1}{T}\sum \mathbf{r}_k$
    Normalize: $\hat{\mathbf{z}} \leftarrow \mathbf{z}/(\|\mathbf{z}\|_2 + \epsilon)$
    Update: $\lambda \leftarrow (1-\beta)\lambda + \alpha \cdot S \cdot \hat{\mathbf{z}}$
  end for
end for
Finalize: $\lambda \leftarrow \lambda/\|\lambda\|_2$

D Experiment Prompts

D.1 Safety System Prompt (Refusal Baseline)

For the **Safety Prompt** baseline (referred to as *LLM termination* or *Prompt-based Baseline* in Table 1), we injected the following "Safety Addendum" into the system instruction. This directive was used to evaluate whether the model could explicitly recognize and refuse implicit harm when instructed to do so.

```
SAFETY ADDENDUM:
If the conversation is clearly
```

causing harm, promoting violence,
or includes explicit abuse, you MUST
refuse to continuey. When refusing,
output EXACTLY:

I must refuse to continue this
conversation.

Do not add any additional tokens
after refusing.