

Content Moderation for Detecting Contextualized Harmful and Hate Speech

Sherine Chally
UCLA
sherinechally@ucla.edu

Arjen Singh
UCLA
asingh0504@ucla.edu

Princeton Liu
UCLA
princetonliu@ucla.edu

Jason Cheng
UCLA
jasonc224@ucla.edu

Brian Yang
UCLA
bryang6465@ucla.edu

Abstract

Hate speech and harmful content in online conversations remain difficult to moderate because toxicity often escalates gradually over multiple turns rather than appearing in a single isolated message. We present a dual model, inference-time content moderation framework that pairs a conversational chatbot with a dedicated safety supervisor. The supervisor monitors a rolling window of conversation history to detect harmful escalation patterns before each chatbot response, then triggers a structured two-phase intervention pipeline when harmful content is detected. Github Source Code: https://github.com/sherinehannahchally/CS263_FinalProject_ContentModeration

1 Motivation

Hate speech moderation is one of the most consequential and technically challenging problems in natural language processing. Platforms that fail to moderate harmful content risk normalizing discrimination, inciting real-world violent, and amplifying radicalization at scale. at the same time, over-moderation that falsely flags benign content erodes user trust disproportionately silences marginalized communities whose legitimate speech may superficially resemble adversarial content.

A central difficulty is that hateful intent is rarely expressed in a single utterance. In practice, users often begin a conversation with innocuous statements, then gradually introduce more extreme or harmful rhetoric over multiple turns—a pattern we call *conversation escalation*. This makes single-turn classifiers ill-suited for real-world moderation, as they lack the temporal context needed to recognize that a sequence of benign-looking messages may collectively constitute a harmful interaction.

We address this gap with a context-aware, inference-time moderation system. Our framework pairs a standard instruction-tuned chatbot with a

dedicated supervisor that evaluates each user message within the full conversation history. We evaluate four supervisor variants—two LLM scales (1B and 8B), a multi-shot prompting strategy, and a fine-tuned BERT baseline—on a synthetic dataset of 400 multi-turn conversations designed to test escalation detection.

2 Related Work

Model-level Safety Alignment The dominant paradigm for safe LLM deployment is to encode safety constraints directly into model parameters through SFT and RLHF (Ouyang et al., 2022). While effective, this approach requires extensive labeled data of harmful interactions and must be repeated whenever new attack categories emerge. Our system operates entirely at inference time, requiring no retraining.

Inference-Time Guardrails. Luo et al. (2025) introduce AGrail, a lifelong agent guardrail that dynamically adapts its safety detection to evolving conversation contexts rather than applying static rules. AGrail is the closest conceptual ancestor to our supervisor architecture: both treat the guardrail as a continuously-running monitor rather than a one-shot classifier. Unlike AGrail, which maintains persistent safety state across sessions, our supervisor operates within a bounded rolling window of recent turns—a design choice that keeps inference cost low while still capturing local escalation dynamics.

Adversarial Training Data Synthesis. Liu et al. (2025) propose IMAGINE, which synthesizes jailbreak-like instructions through iterative latent space expansion to pre-train more robust rejection boundaries. Our work similarly relies on synthetically generated harmful conversations as evaluation data, but uses a generative pipeline based on an ablated LLaMA model rather than latent space

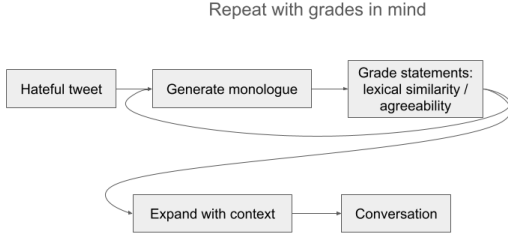


Figure 1: Conversation generation flow

manipulation. Critically, we use synthetic data only for evaluation, not for retraining the supervisor.

Novel Attack Vectors. Cisneros-Velarde (2025) demonstrate that humor-based prompts can bypass safety guardrails by exploiting distributional gaps in alignment training data. This work motivates our multi-shot prompting strategy: by exposing the supervisor to diverse harmful styles (including subtle, indirect hate speech) in its context window, we aim to reduce reliance on surface-level keyword matching that humor-based attacks exploit.

3 Data

3.1 Source Data

We draw seed tweets from the hate speech and offensive language dataset (Davidson et al., 2017), which labels tweets as hate speech (class 0), offensive language (class 1), or neither (class 2). We exclude the ambiguous offensive class entirely and sample 200 hate speech seeds and 200 neutral seeds, yielding a balanced binary split.

3.2 Synthetic Conversation Generation

All conversations are generated by a Llama-3.3-8B-Instruct model with weakened refusal mechanisms (“abliterated”), loaded in 4-bit quantization. This model is used because standard instruction-tuned models refuse to produce the gradually escalating harmful content needed for adversarial evaluation.

Adversarial Conversations. For each hate speech seed tweet, the generator model produces a multi-turn monologue that begins with benign statements and escalates toward the seed’s sentiment, using dogwhistles, thought-terminating cliches, and vague language to do so. Each draft is then refined over two iterative rounds.

1. **Similarity Grading.** The 8B generator model scores each line’s agreement in comparison to the seed tweet on a scale of 1 to 10.

2. **Bypass Scoring.** The 1B supervisor classifies the seed tweet alone and then the full conversation prefix. Lines where the prefix causes a BLOCK→PASS flip are marked as effective; others signal the revision loop to push harder.

3. **Revision.** The generator rewrites the monologue conditioned on both scores, aiming for low early similarity and high late similarity.

Subsequently, the list of scores was used to iteratively revise the monologues before they get fed into the supervisor model.

Benign Conversations. Neutral seeds undergo the identical pipeline: initial draft, iterative similarity grading, revision. We instruct the generator to produce neutral, topically relevant buildup. All turns, including the seed are labeled 0. This ensures benign conversations are structurally indistinguishable from adversarial ones (multi-turn, LLM-generated, comparable length distribution), preventing supervisors from exploiting surface-level formatting cues.

3.3 Conversation Generation Results

The final generated dataset contains 400 conversations totaling 7,494 turns (mean 18.7 turns/conversation). Adversarial conversations average 20.3 turns (range 9–167) and benign conversations average 17.1 turns (range 7–86). The mean context length ratio between adversarial and benign conversations is $1.65\times$, reflecting the tendency of escalation sequences to produce longer outputs during revision. We note this imbalance as a confound: supervisors may partially learn to associate longer contexts with higher risk. Only 2.7% of turns carry the positive label (200 out of 7,494), making turn-level precision uninformative; we therefore evaluate primarily at the conversation level.

4 Methods

4.1 System Architecture

All configurations in our framework share the same structural principle: a **generator** model paired with a **supervisor** model at inference time. In deployment, the supervisor evaluates each user message within the full conversation context, then the generator responds; if the supervisor issues a BLOCK verdict, the system activates an intervention rather than generating a reply. The two models share no

existing internal state and do not exchange activations.

4.2 Supervisor Variants

We use four supervisors spanning two architectures and three prompting strategies. All supervisors receive the *full conversation prefix* up to the current turn so that escalation patterns across turns are visible.

Zero-Shot 1B. A 4-bit quantized Llama-3.2-1B-Instruct (Q4_K_M GGUF, context window 4096 tokens); The supervisor receives a system prompt that instructs it to classify user intent without answering the query. The zero-shot prompt defines a minimal rule set: block if the text requests illegal activity, violence, hacking, or harm; pass otherwise. No examples are provided. This configuration establishes a baseline for how well a small LLM can moderate content through instruction alone.

Multi-Shot 1B. The multi-shot configuration extends the zero-shot prompt by prepending four labeled examples covering distinct hate speech categories: racism, sexism, homophobia, and xenophobia. The system prompt also enumerates explicit BLOCK criteria (racism, sexism, homophobia, violence, illegal activity). By grounding the supervisor with in-context demonstrations, multi-shot prompting is intended to improve recall on subtle or indirect harmful content.

Zero-Shot 8B. A 4-bit quantized Llama-3.3-8B-Instruct-abliterated (Q4_K_M GGUF, context window 4096 tokens) with the same zero-shot prompt as the 1B variant. This model also serves as the generator during conversation generation (Section ??); reusing it as a supervisor tests whether a larger model provides better classification at higher inference cost. Because the same model generates the adversarial conversations and later judges them, we note this dual role as a potential confound.

HateBERT. A fine-tuned BERT encoder (dehatebert-mono-english) representing the pre-LLM supervised classification baseline. Unlike the LLM supervisors, HateBERT was trained directly on hate speech labels with a classification head and requires no prompt engineering. It processes each conversation prefix as a flat text input (truncated to 512 tokens) and outputs $P(\text{HATE})$; we apply a threshold of 0.5 to produce a binary verdict. At ~ 5 ms per call, it is orders of

magnitude faster than the generation-based LLM supervisors.

Verdict Extraction. The three LLM supervisors generate a short response beginning with VERDICT: BLOCK or VERDICT: PASS. Generation is halted after the first newline (~ 4 tokens) via a stop sequence, since only the verdict line is used for evaluation. A regex parser extracts the verdict; any output not matching PASS is treated as BLOCK.

Logit-Only Classification (Negative Result).

We additionally tested single-token logit extraction: reading the raw logit for BLOCK vs. PASS at the first predicted token position after the prompt, without generating any tokens. All three logit variants produced degenerate all-BLOCK classifiers (recall = 1.0, precision = 0.027) regardless of input content, indicating that the instruction-tuning safety prior at this token position overwhelms per-input signal. Generation-based decoding, which gives the model multiple tokens to deliberate, is necessary for meaningful classification. We report this as a negative result and exclude the logit variants from further analysis.

4.3 Context-Aware Evaluation

At each turn t in a conversation, the supervisor receives the concatenation of all turns from turn 0 through turn t , separated by newlines:

$$\text{context}_t = u_0 \parallel u_1 \parallel \dots \parallel u_t$$

This full-prefix design allows the supervisor to observe gradual escalation, rather than evaluating each message in isolation. The cost is linear prompt growth; the longest conversation in our dataset (167 turns) approaches the 4,096-token context window.

4.4 Parallel Inference

To reduce evaluation wall time, we load two independent copies of the 1B supervisor model (~ 800 MB VRAM each), each guarded by a separate mutex. The zero-shot and multi-shot supervisors are assigned to different instances and run concurrently via a thread pool, halving the 1B generation wall time. The 8B supervisor runs sequentially on the shared generator model.

Dual History Design. A key design decision is the maintenance of two separate conversation histories:

1. `chat_history`: contains only turns the supervisor has cleared as safe. This is the context passed to the chatbot for response generation, ensuring the chatbot is never exposed to harmful content through its own context window.
2. `supervisor_history`: contains *all* turns, including blocked ones. Used exclusively by the supervisor, this complete record enables detection of gradual escalation patterns even after a prior intervention.

This separation guarantees that a blocked turn cannot influence the chatbot’s future behavior while still allowing the supervisor to reason about the full conversational trajectory.

5 Results

We presented a dual-model content moderation framework that performs context-aware, inference-time supervisor without model retraining. Table 1 summarizes evaluation results for the four working supervisors.

	zero_shot	multi_shot	8B	HateBERT
Convo F1	0.706	0.717	0.686	0.799
Convo Precision	0.632	0.620	0.524	0.905
Convo Recall	0.800	0.850	0.995	0.715
Early Det. Rate	0.385	0.375	0.810	0.235
Miss Rate (final)	0.220	0.155	0.005	0.285
Bypass Rate	0.090	0.130	0.000	0.220
Gradient ρ	0.546	0.542	0.702	0.510
FP Count	1,622	904	3,269	411

Table 1: Conversation-level evaluation across four supervisor variants. **Bold** indicates best per metric. Early detection rate = fraction of adversarial conversations flagged before the final harmful turn. Bypass rate = fraction where adversarial wrapping flips BLOCK→PASS (white-box for zero_shot; transfer for others). Gradient ρ = Spearman correlation between conversation position and block rate. FP count is turn-level (out of 7,294 benign turns).

5.1 Precision, Recall, and Latency

The four supervisors occupy distinct points in a tradeoff between precision, recall, and detection latency.

HateBERT achieves the highest conversation-level F1 (0.799) and precision (0.905), but detects only 23.5% of adversarial conversations before the final harmful turn and misses 28.5% entirely. This pattern is consistent with a classifier that recognizes explicit toxicity but lacks the discourse-level reasoning to detect gradual escalation.

The 8B model occupies the opposite extreme: 99.5% conversation recall and 81% early detection, missing only a single adversarial conversation. However, its 52.4% precision means it flags roughly half of all benign conversations as harmful. In a deployment setting, this over-triggering would render the system unusable.

The two 1B LLM supervisors fall between these extremes. Multi-shot achieves the best balance among LLM variants: 85% recall with 62% precision, yielding a conversation F1 of 0.717. The four in-context exemplars improve recall by 5 percentage points over zero-shot (0.850 vs. 0.800) while raising precision from 63.2% to 62.0%—a negligible tradeoff. Zero-shot performs comparably but with slightly more conservative blocking (63.2% precision, 80% recall).

5.2 Adversarial Robustness

The adversarial conversations were optimized to bypass the zero-shot 1B supervisor during conversation generation; its 9% bypass rate is therefore a white-box measure. All other bypass rates are transfer rates, measuring generalization to unseen supervisors.

HateBERT is $2.4\times$ more bypassable (22%) than the optimization target, confirming that encoder-only classifiers are more vulnerable to adversarial escalation than LLM-based supervisors that can reason over discourse structure. Multi-shot 1B transfers at 13%, slightly above the white-box rate, suggesting that the few-shot exemplars anchor the model’s expectations and create a marginally wider attack surface. The 8B model’s 0% bypass rate reflects its near-universal blocking rather than genuine robustness.

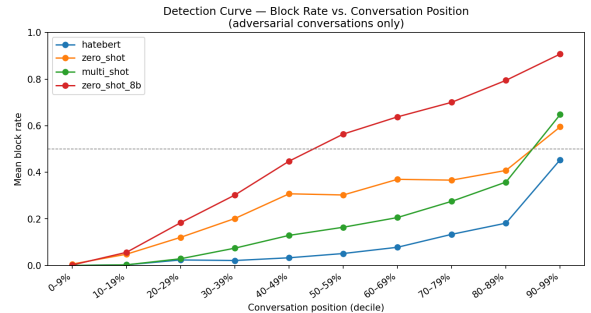


Figure 2: Mean block rate by conversation position decile (adversarial conversations only). Rising curves confirm supervisors detect escalation rather than firing randomly. The 8B model exceeds 0.5 by the 40% mark; HateBERT remains below 0.5 until the final decile.

6 Conclusion

We presented an inference-time content moderation framework that evaluates each user message within the full conversation context, enabling detection of gradually escalating harmful content that single-turn classifiers miss. Our evaluation across four supervisor variants on 400 synthetic conversations reveals three key findings.

First, context-aware LLM supervisors can detect adversarial escalation before explicit harmful content appears. The 8B model flags 81% of adversarial conversations early, and even the smaller 1B variants detect roughly 38% early turns before any labeled-harmful content is present. This demonstrates that passing conversation history to the supervisor provides a qualitatively different capability than isolated turn classification.

Second, no single supervisor dominates across all metrics. HateBERT provides the highest precision (0.905) and F1 (0.799) but is $2.4\times$ more vulnerable to adversarial bypass than the LLM supervisors and rarely detects escalation early. The 8B model catches nearly everything but blocks half of all benign conversations. The 1B multi-shot variant offers the most practical balance between safety and usability. This tradeoff suggests that production systems should consider ensemble or cascading architectures that combine a high-precision first stage with a high-recall second stage.

Limitations. Our evaluation has several limitations. The adversarial conversations were optimized against one supervisor (zero-shot 1B), making bypass rates for other variants transfer rates rather than direct robustness measures. The $1.65\times$ context length imbalance between adversarial and benign conversations is a potential confound. All buildup turns are labeled 0, though late-stage turns likely contain genuinely harmful content—our label quality audit estimates this noise at approximately the rate flagged by the 8B model in isolation. Finally, we evaluate supervisors on synthetic conversations; generalization to organic multi-turn discourse remains untested.

Future Work. Natural extensions include evaluating on organic conversation data, testing cascading supervisor architectures that route easy cases through HateBERT and escalate uncertain cases to an LLM supervisor, and investigating whether KV-cache reuse across turns within a conversation can reduce the linear prompt growth cost of full-

prefix evaluation. Future deployments could replace blunt refusals with a two-phase intervention that acknowledges the flagged content and offers educational resources before terminating the session.

Future efforts could also be made toward the interpretability of content moderation models. Although large models such as HateBERT and LLM-based supervisors can achieve strong performance, their inference-time behavior remains largely opaque beyond observable outputs. Techniques such as attention visualization, probing (Liu et al., 2024), and feature attribution (Taimeskhanov and Garreau, 2025) could help determine whether the model is relying on semantic signals or spurious correlations.

Lastly, human-focused evaluation frameworks and metrics should be established. This includes defining metrics for explanation usefulness, robustness to adversarial inputs, or conducting human studies to assess whether explanations genuinely improve user trust and supervisor effectiveness. Advancing interpretability alongside model-alignment would enable future moderation systems to become not only more accurate, but also more accountable and better suited to real-world deployment.

Contribution Statement

- **Brian Yang & Jason Cheng:** Designed and implemented the conversation generation pipeline, generating adversarial multi-turn conversations from seed tweets via iterative monologue drafting, similarity grading, rewriting with context, and language guidelines for diverse harmful rhetoric.
- **Princeton Liu & Sherine Chally:** Designed and implemented the zero-shot, multi-shot, and multi-model ensemble supervisor configurations, the full supervisor-chatbot architecture, dual history design, context-aware prompting.
- **Arjen Singh:** Implemented the full evaluation framework, including turn-level and conversation-level metrics, detection latency computation, synthetic data quality analysis using TTR and Distinct- n statistics, HateBERT integration, and the two-phase intervention pipeline.

All team members contributed to writing and revising the paper.

References

- Pedro Cisneros-Velarde. 2025. Bypassing safety guardrails in llms using humor. *arXiv preprint arXiv:2504.06577*.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.
- Sheng Liu, Qiang Sheng, Danding Wang, Yang Li, Guang Yang, and Juan Cao. 2025. Forewarned is forearmed: Pre-synthesizing jailbreak-like instructions to enhance llm safety guardrail to potential attacks. *arXiv preprint arXiv:2508.20038*.
- Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Haonan Lu, Bing Liu, and Wenliang Chen. 2024. Probing language models for pre-training data detection. *arXiv preprint arXiv:2406.01333*.
- Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, and Chaowei Xiao. 2025. Agrail: A lifelong agent guardrail with effective and adaptive safety detection. *arXiv preprint arXiv:2502.11448*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Magamed Taimeskhanov and Damien Garreau. 2025. Feature attribution from first principles. *arXiv preprint arXiv:2505.24729*.