# Scalable Detection of Adversarial Synthetic Slop and Coordinated Media Abuse: A LoRA-Enabled Multimodal Defense System

Abhinav Mathur, Claire Liu, Kelvin Tan, and Yifei Liu
*Google*
{abhinavmathur, lclaire, weijiekelvin, yifeili}@google.com

***Abstract*—Online video platforms face an exponential challenge in detecting and mitigating the flood of AI-generated "slop" and synthetic spam perpetuated by coordinated malicious actors. This content is increasingly designed to exploit the limitations of traditional media forensics, often utilizing generative AI to produce unique, localized variations of harmful or low-quality material at scale. Traditional content-centric moderation fails against this coordinated, adversarial generation strategy.**

**This paper presents a novel, scalable defense system designed for online video platforms (OVP) to identify and terminate clusters of coordinated accounts exhibiting a prevalence of adversarial synthetic content. The approach leverages a multi-faceted architecture incorporating two core machine learning components: a robust Coordinated Bot-Net Detector (via Account Relatedness) and a Synthetic Pattern Classifier. Crucially, we introduce an advanced AI enhancement layer utilizing Large Language Models (LLMs), specialized via Low-Rank Adaptation (LoRA) and Automatic Prompt Optimization (APO), to achieve rapid, high-precision semantic understanding of emerging synthetic spam trends.**

**Test data demonstrates the system's significant impact, resulting in the successful termination of clusters at a high precision comprising channels of synthetic spam generators. Furthermore, the LLM-driven automation significantly improves operational efficiency, resulting in significant human review efficiency gains. This work details a critical system design that provides essential scalability and adversarial resilience against sophisticated generative attacks.**

## I. Introduction

Online video platforms are constantly targeted by malicious actors leveraging **Generative AI tools** to exploit platform scale for illicit purposes. A significant challenge arises from coordinated groups who utilize these tools to flood platforms with **"AI slop"**—content that is mass-produced, often borderline in nature, and designed to overwhelm quality filters. These actors engage in **adversarial adaptation**, continuously modifying synthetic outputs to strategically stay below violation thresholds while redirecting users to off-platform scams or harmful services.

The scale and speed of this **Generative AI spam** make it difficult to detect using traditional hashing or metadata-based approaches, as the generative process creates unique fingerprints for functionally identical content. This manuscript details a comprehensive engineering system evaluated to address this **synthetic media scalability challenge**.

The system moves beyond isolated content analysis by focusing on identifying **bot-net clusters** exhibiting synchronized behavior and a prevalence of synthetic artifacts. We define this content not just by policy, but by **Content Integrity Standards**, targeting verticals highly susceptible to generative abuse such as synthetic impersonation, procedural gore, and AI-generated scams.

The work presented here describes the **Scalable Cluster Termination System (S-CTS)**, which identifies coordinated account clusters with a certain percentage of channels exhibiting demonstrable synthetic patterns. Furthermore, the LLM integration layer addresses the complexity of detecting these subtle, often "realistic" AI generations through a two-stage multimodal classification architecture.

## II. Related Work

The S-CTS system integrates engineering paradigms across three distinct but related domains: Sybil attack detection, synthetic media forensics, and the application of LLMs in adversarial security.

### A. Detecting Account Relatedness and Bot-Nets

Identifying coordinated synthetic actors often begins with linking seemingly disparate accounts (Sybil detection). Existing techniques leverage various features to establish account relatedness, including shared IP addresses and device identifiers. **Seminal works such as SybilGuard [1], SybilLimit [2], and SybilRank [3] have demonstrated the efficacy of graph-based structural analysis in identifying coordinated groups of fake identities.**. Our work leverages internal algorithms that analyze a combination of infrastructure-level signals and inorganic behavioral patterns to establish **"Generation Clusters"**—groups of accounts likely utilizing the same generative API or script.

### B. Synthetic Media Forensics and Similarity

Effective detection of "AI Slop" requires robust similarity analysis that goes beyond exact matching. Traditional content-centric moderation paradigms struggle in this space because they treat trust and safety as an aggregation of isolated, individual post-by-post decisions [6]. This structural vulnerability is heavily exploited by adversarial networks, which utilize generative AI to produce infinite, unique variations of functionally identical spam, bypassing traditional cryptographic hashing and simple metadata filters [8]. By shifting the defense

vector from individual content evaluation to systemic account-relatedness and behavioral clustering, the S-CTS system addresses the root organizational structure of these campaigns rather than their downstream outputs [7].

For text-based content, methods like text embeddings generated by models like Sentence-BERT are used to detect scripted AI narratives. For multimedia, traditional techniques include perceptual hashing. However, generative AI introduces unique challenges; our system employs proprietary algorithms that analyze both textual and multimedia content to identify **"Generative Artifacts"**—subtle markers of synthetic production shared across channels.

While the industry is coalescing around cryptographic provenance standards defined by the **Coalition for Content Provenance and Authenticity (C2PA)** and imperceptible digital watermarking technologies (e.g., Google DeepMind's SynthID) as the "gold standard" for establishing media authenticity, adversarial adoption remains a critical gap. Malicious actors frequently utilize open-source models lacking these safeguards or actively strip provenance metadata. Consequently, until such affirmative authenticity signals become ubiquitous and tamper-proof across the open web, scalable **detection systems** like S-CTS remain the primary defense mechanism against non-compliant, coordinated synthetic abuse.

### C. Large Language Models in Security

The application of LLMs for security is a rapidly evolving field. While some work discusses the challenges of classifying ambiguous content, our work directly addresses the engineering gap of latency and cost in utilizing these models at scale. By utilizing LoRA and APO, we demonstrate a practical framework for "Adversarial Ops"—using AI to catch AI—significantly enhancing the efficiency and precision of the detection pipeline against shifting generative models.

## III. System Design and Architecture

Our system for detecting coordinated synthetic abuse comprises several key stages:

### A. Classifier $\Psi_A$: Coordinated Bot-Net Detection

The $\Psi_A$ component utilizes internal Google algorithms to analyze proprietary infrastructure signals. This uses proprietary signals that Google has which are very similar to known behaviors like API usage patterns, event time series analysis, and GenAI-specific metadata. We leverage algorithms to cluster content and find sub-clusters which show common inorganic behavior or vice versa. This allows us to identify high-confidence clusters of accounts that are statistically likely to be controlled by the same actor or automated generation script.

### B. Classifier $\Psi_C$: Synthetic Content & Slop Prevalence

The $\Psi_C$ component is responsible for scoring content against specific **Content Integrity benchmarks**. It employs proprietary algorithms that analyze textual (e.g., text embeddings for AI-generated scripts) and multimedia components

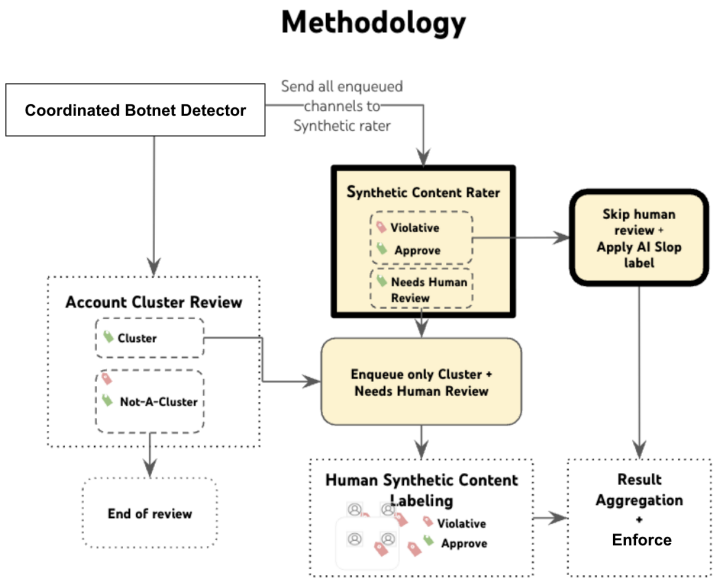


Fig. 1. The S-CTS Engineering Workflow. The system integrates cluster detection signals with the LLM-based content rater. Channels with high-confidence LLM scores bypass human review, while ambiguous cases are routed to human experts, ensuring high precision.

(e.g., deep feature extraction for visual artifacts) to provide content-level synthetic classification tags.

### C. Engineering Workflow

Potential malicious clusters are identified at the intersection of high $\Psi_A$ confidence scores and the presence of similar synthetic-tagged content identified by $\Psi_C$. **It's important to highlight the latency advantage that we gain by grouping accounts which reduces the compute cost per decision compared to individual video scans**.

## IV. Advanced Detection: LLMs and Parameter-Efficient Fine-Tuning (LoRA)

The integration of advanced Large Language Models (LLMs) represents a significant engineering enhancement to the $\Psi_C$ component, dramatically improving the scalability and precision of **AI slop detection**.

### A. Two-Stage LLM Architecture for Synthetic Artifact Extraction

To effectively manage the high dimensionality of video data, we developed the **Synthetic Content Rater**, a two-stage architecture designed to transform raw multimodal signals into actionable textual context.

- **Stage 1: Multimodal Context Distillation (Feature Extraction):** This stage processes the raw video context, which includes video frames, audio tracks, and transcripts. Rather than relying on pixel-level forensic analysis (e.g., visual inconsistencies), this stage leverages **multimodal semantic embeddings** and **temporal behavioral signals** to identify automated content generation.

Specifically, the model analyzes **Video Text Embeddings** (`Feature_video_text_embedding`) and **Salient Terms** (`Feature_title/desc_salient_terms`) to detect repetitive, templated narratives common in AI-generated "slop". Simultaneously, it evaluates **Upload Pacing** (`Feature_avg_log_upload_pace`, `Feature_time_to_first_upload_secs`) to identify non-human, high-frequency publishing behaviors characteristic of automated scripts. These signals are combined with **Pulsar visual embeddings** to categorize the semantic nature of the content. It produces a compact, integrity-aware textual summary, significantly reducing the downstream processing load.

- **Stage 2: Channel-level Classifier:** This stage makes the final determination, leveraging the textual context summaries. This design addresses the limitation of traditional classifiers by focusing the core LLM task on **semantic reasoning over synthesized features** rather than raw video pixels.

### *B. LoRA and APO Adaptation for Adversarial Resilience*

The Stage 2 Classifier is specialized for synthetic trend detection using **Parameter-Efficient Fine-Tuning (PEFT)** techniques, specifically **Low-Rank Adaptation (LoRA)** and **Automatic Prompt Optimization (APO)**.

- **Engineering for Scale:** This approach allows for the efficient adaptation of the large proprietary LLM (e.g., Gemini 2.0 Flash) without the prohibitive computational cost of full fine-tuning. Specifically, LoRA significantly reduces the number of trainable parameters and **substantially decreases** the memory footprint, allowing for rapid, cost-effective execution and parallelized inference on scalable TPU infrastructure.
- **Rapid Adaptation:** APO allows us to engineer prompts that adapt to new "Slop" trends faster than retraining a dense model. We can retrain a LoRA adapter **rapidly** when a new GenAI model (like Sora or Kling) is released by attackers.

### *C. Automated Enforcement and Scalability*

The semantic scores generated by the LoRA-enhanced LLM drive automated decisions.

- **Automate VIOLATES ($\tau_V$):** When the synthetic likelihood score exceeds a high threshold, the system triggers auto-enforcement.
- **Automate APPROVES ($\tau_A$):** High recall for automated approval ensures scalability by shunting the vast majority of authentic content from the pipeline.

This architecture is designed to be highly scalable, processing data in batches. To ensure high precision, the pipeline operates with a fixed temporal aggregation window. This batching window allows sufficient time to gather necessary multimodal signals before triggering automated classification, significantly reducing the manual review bottleneck.

## V. Evaluation and Empirical Results

The effectiveness of S-CTS is evaluated based on system throughput, efficiency metrics, and the quantified performance of the LLM-enhanced classifiers on live streams of synthetic media.

### *A. Operational Impact (6-Month Baseline)*

The system successfully mitigated coordinated synthetic abuse:

- **Cluster Validation Turnaround Time (Avg.):** 32% reduction compared to humans
- **Synthetic Content Review Turnaround Time (Avg.):** 50% reduction compared to humans

The extremely low **False Positive Rate** validates the high operational accuracy and precision of the system when distinguishing between **"Creative AI Use"** and **"Adversarial Slop."**

### *B. LLM-Enhanced Classification Performance on Synthetic Datasets*

The following table presents the performance of the LLM-enhanced classification component ($\Psi_C$) across several challenging synthetic verticals.

### *C. Analysis of LLM Automation and Precision Trade-offs*

For automated enforcement, thresholds are consistently set high to yield high precision (92% to 95%). This design choice prioritizes minimizing false positives, essential when applying classification decisions **against AI-generated content** to avoid censoring legitimate creators using AI tools. This design also incorporates a protection mechanism designed to prevent inadvertent takedowns. Conversely, high recall (up to 96%) for automated approval decisions ensures scalability by shunting the vast majority of benign content from the review process.

At the time of writing, empirical results for the latest generative models (e.g., Sora, Kling) are limited by the scarcity of large-scale, ground-truth adversarial datasets. However, our architectural hypothesis rests on the agility of **APO and LoRA**, which require **orders of magnitude fewer** labels compared to retraining dense classifiers. While base foundation models may have inherent limitations in distinguishing synthetic distributions due to their training data, our framework utilizes the LLM as a **semantic reasoner** over curated, trend-specific datasets. To address current data gaps, our proposed pipeline is designed to ingest signals from specialized external synthetic-media classifiers as supplementary features, ensuring the system evolves in lock-step with emerging generative trends.

## VI. Ethical Considerations and Responsible AI Engineering

### *A. Distinguishing "Slop" from Creativity*

Synthetic media classification is highly susceptible to definition drift. To mitigate the risk of algorithmic over-enforcement against **legitimate AI artists**, the system enforces a precision-over-recall mandate. Crucially, the "Cluster" requirement acts

TABLE I
PERFORMANCE METRICS FOR LLM-ENHANCED SYNTHETIC MEDIA CLASSIFICATION (LORA TRAINING RESULTS)

| Synthetic Content Category | Decision Type | Precision | Recall |
|---|---|---|---|
| **Synthetic Text-to-Speech Narratives** (e.g., Racy Stories) | Automate VIOLATES | 95% | 83% |
| **Synthetic Text-to-Speech Narratives** | Automate APPROVES | 93% | 95% |
| **Generative NSFW/Slop** (Sex & Nudity) | Automate APPROVES | 88% | 95% |
| **Generative NSFW/Slop** | Automate VIOLATES | 95% | 68% |
| **Procedural Shock/Gore** (Child Safety) | Automate APPROVES | 72% | 95% |

as a safeguard: we primarily target **coordinated, mass-produced behaviors** rather than isolated uploads, reducing the risk of penalizing individual creators experimenting with new tools.

### B. Fairness and Robustness of LoRA Fine-Tuning

While computationally efficient, the use of LoRA fine-tuning necessitates careful evaluation. Low-rank fine-tuning can inadvertently preserve undesirable biases embedded within the massive foundation model. LLM decisions are subject to a **periodic expiration policy** to **prevent enforcement based on outdated data**. The LoRA adaptation process must be rigorously monitored to ensure it captures the necessary semantic shift for **Synthetic Integrity** without amplifying biases.

## VII. CONCLUSION AND FUTURE WORK

### A. Conclusion

The S-CTS system represents a scalable and highly accurate defense against coordinated actors exploiting Generative AI to flood platforms. The success hinges on the robust integration of multi-account coordination detection (Bot-Nets) and the utilization of LoRA-adapted LLMs for nuanced synthetic content classification. Operational evaluations confirm the system's impact on identified coordinated clusters, and the exceptionally low $< 1\%$ overturn rate validates the approach's engineering resilience.

### B. Future Work

Ongoing research efforts focus on continuous engineering improvements:

- **Integration with C2PA and Watermarking:** Future work will ingest cryptographic signals (C2PA) and digital watermarks (e.g., SynthID) as ground-truth features for the $\Psi_C$ classifier, moving from "detection" to "provenance verification."
- **Expanding to Deepfake Impersonation:** The LLM-driven detection framework will be extended to specifically target high-harm deepfakes involving political figures or non-consensual imagery.
- **Tracking Adversarial Adaptation:** The team will continue to leverage LLMs for daily monitoring to track emerging "Slop" trends, ensuring rapid adaptation of detection models to new foundational models released by the open-source community.


## ACKNOWLEDGMENTS

The authors would like to thank the engineering team, the operations team, and the policy development team for their invaluable input and operational support in developing and evaluating the Scalable Cluster Termination System.


## REFERENCES


[1] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "SybilGuard: defending against sybil attacks via social networks," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '06)*, pp. 267–278, https://dl.acm.org/doi/10.1145/1159913.1159945

[2] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks," in *2008 IEEE Symposium on Security and Privacy*, pp. 3-17, https://ieeexplore.ieee.org/document/4531145

[3] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12)*, pp. 197–210, https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/cao

[4] "Ethics in Fine-Tuning AI Models: Bias, Responsibility, and Compliance," GoCodeo, accessed October 27, 2025, https://www.gocodeo.com/post/ethics-in-fine-tuning-ai-models-bias-responsibility-and-compliance

[5] "Low-rank finetuning for LLMs: A fairness perspective," ResearchGate, accessed October 27, 2025, https://www.researchgate.net/publication/380974194_Low-rank_finetuning_for_LLMs_A_fairness_perspective

[6] **Douek, E. (2021).** *Governing Online Speech: From "Individual Adjudications" to "Systemic Design."* Journal of Free Speech Law, 1(1), https://harvardlawreview.org/wp-content/uploads/2023/02/136-Harv.-L.-Rev.-1030-Fox.pdf

[7] **François, C., & Douek, E. (2021).** *Tactical Evolution in Coordinated Inauthentic Behavior: How Adversarial Networks Evade Content-Centric Classifiers.* Journal of Online Trust and Safety, 2(1). https://doi.org/10.54501/jots.v1i1.17

[8] **Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023).** *Universal and Transferable Adversarial Attacks on Aligned Language Models.* arXiv preprint arXiv:2307.15043, https://arxiv.org/abs/2307.15043