The Synthetic Gap: Automating Forensic Investigation of "AI Slop" with the Scaled Abuse Forensics Examiner (SAFE)

Abhinav Mathur*, Crystal Zhao*, Geethik Narayana Kamineni*, Longling Wang*,
Lucas Liu*, Utkarsh Chaudhary*, and Vahid Jalali*
*Google
{abhinavmathur, crystalzhao, geethik, llw, lucasliu, utkarshchdhry, vjalali}@google.com

***Abstract*—Generative AI capabilities have enabled malicious actors to flood online platforms with "AI slop"—mass-produced, low-quality synthetic media designed to overwhelm traditional integrity systems. These adversarial campaigns often utilize coordinated networks to distribute unique, localized variations of synthetic content, rendering static detection methods ineffective. The signals to detect coordination often have recall gaps. The content is not exactly duplicative to be in the same repetitive video cluster. The abusers however show similar patterns of behavior which need forensics. Manual forensic investigations cannot scale to match the velocity of these generative attacks.**

To address this, we present SAFE (Scaled Abuse Forensics Examiner), an automated multi-agent architecture designed for the scalable forensics of adversarial synthetic media. The system decomposes the investigation process into specialized agents: a Cluster Understanding Agent specialized in analyzing the relations between channels in a cluster, a Behavior Understanding Agent that identifies inorganic spatiotemporal patterns, and a Content Understanding Agent that utilizes LoRA-adapted Large Language Models (LLMs) and few-shot learning to detect existing policy violations and spirit of the policy violations respectively. A Root Agent synthesizes these multimodal signals to render a final verdict. Early deployment results indicate that SAFE significantly accelerates the identification of novel synthetic threats, reducing forensic investigation time compared to human-in-the-loop workflows.

***Index Terms*—Generative AI, Multi-Agent Systems, Platform Integrity, Digital Forensics, Policy Enforcement, Deepfake Detection.**

I. Introduction

The proliferation of Generative AI tools has fundamentally altered the landscape of platform integrity. Adversaries now possess the capability to generate and distribute synthetic video content at an unprecedented scale, creating "bot-nets" of channels that systematically upload "AI slop" or borderline synthetic/abusive material. These coordinated networks engage in adversarial adaptation, continuously modifying their generative prompts to produce content that evades static classifiers while maximizing reach.

Traditional forensic workflows, which rely heavily on manual pattern recognition and metadata analysis, are ill-equipped to handle this volume. The "synthetic gap"—the time between the emergence of a new generative attack vector and the deployment of a counter-measure remains a critical vulnerability.

This paper introduces **SAFE**, a novel engineering solution that leverages a Multi-Agent System (MAS) to automate the forensic investigation of suspicious clusters. Unlike traditional classifiers that output a simple probability score, SAFE mimics the reasoning process of a forensic analyst. It autonomously retrieves and correlates distinct pillars of evidence—infrastructure signals, inorganic behavioral patterns, and generative content artifacts to identify and explain coordinated synthetic abuse.

II. Background

The literature underpinning the SAFE architecture spans three critical areas: the detection of inorganic adversarial activity, scalable multi-agent systems for digital forensics, and the application of transformer-based architectures for semantic content understanding and policy violation detection.

A. Detecting Inorganic Behavior

The proliferation of bot-nets and coordinated adversarial campaigns necessitates robust methods for identifying non-human engagement patterns. Research has focused on analyzing entity clusters across platforms to expose manipulated metrics and misinformation. Techniques such as rolling window correlation and k-means clustering applied to temporal and semantic features have proven effective in classifying accounts as organic or inorganic. For example, studies have assessed thousands of YouTube channels to find "inorganic" activity used to boost engagement metrics [3], while other work has classified social media accounts based on their temporal and semantic posting features, highlighting strategies like the "follow-refollow" approach [1].

B. Automating Forensics with Multi-Agent Systems

As synthetic content generation accelerates, forensic detection must be automated and scalable. The concept of using specialized AI agents to collaborate on complex tasks has emerged as a promising solution. Frameworks like DeepAgent [4] utilize dual-stream multi-agent fusion, where specialized agents analyze visual artifacts and audio-visual consistency to detect deepfakes, achieving high accuracy in cross-dataset validation. Furthermore, multi-agent systems are leveraged for simulating and detecting malicious activities in enterprise environments, such as the Chimera framework [7] which automates the generation and monitoring of complex system logs to

simulate and detect insider threats. These models demonstrate the power of LLM-driven agents to uncover semantically coherent, yet malicious, "inorganic" activities.

C. Transformer-Based Content Understanding for Policy Enforcement

Beyond simple forgery detection, modern architectures are increasingly leveraged for deep semantic analysis to identify violations of platform safety policies. Recent advancements in Multimodal Large Language Models (MLLMs) and Vision Transformers (ViT) allow for the simultaneous processing of visual and textual cues to detect nuanced harms like hate speech, harassment, or coordinated influence operations. For instance, researchers have utilized contrastive learning and cross-attention mechanisms to identify toxic content that evades traditional keyword filters [5]. Additionally, the use of Retrieval-Augmented Generation (RAG) within transformer frameworks has enabled real-time verification of content against evolving community guidelines, allowing for the detection of policy-violating narratives even when the media itself is technically "organic" but contextually harmful [6]. These high-precision deep learning frameworks shift the focus from pixel-level artifacts to the broader intent and impact of the media.

III. SAFE Architecture: A Multi-Agent Approach

We propose a modular, hierarchical architecture comprising a Root Agent and specialized sub-agents, each tasked with a specific domain of synthetic forensics. The figure below shows a simplified visual representation of our system design.

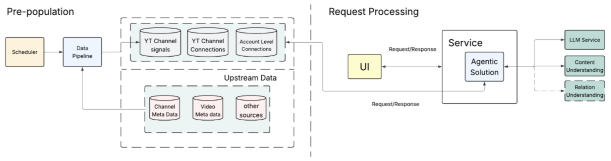

Fig. 1. SAFE system design and high-level structural pipeline.

In this section we zoom into the agentic solution module which is represented in the figure below

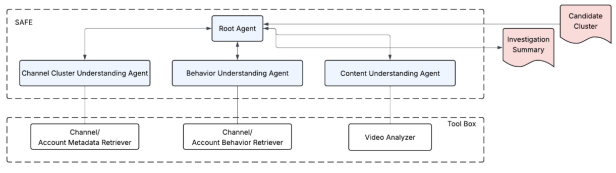

Fig. 2. Detailed schematic of the SAFE multi-agent architecture interaction layer.

A. Root Agent (The Orchestrator)

The Root Agent serves as the central reasoning engine. It accepts a candidate cluster of channels as input and dynamically delegates tasks to sub-agents. It is responsible for sub-agent verdict *judgement* and *context synthesis*: vetting the output generated by the sub-agents and aggregating the structured evidence from the Relationship, Behavior and Content agents to determine if the cluster represents a coordinated synthetic attack or organic activity.

B. Content Understanding Agent (Synthetic Artifact Detection)

The Content Understanding Agent is tasked with differentiating between sophisticated generative media and mass-produced "AI slop," while identifying content that falls into prohibited synthetic categories.

- **Violative content detection**: This module utilizes a Large Language Model (LLM) leveraging Low-Rank Adaptation (LoRA) [2] to achieve specialized detection accuracy.
- **Spirit of policy violation**: To address adversarial adaptation, this component employs a few-shot trained LLM to capture nuanced violations that evade traditional classifiers and the fine-tuned model by aligning with the "spirit" of platform guidelines. In addition, we plan to introduce a policy understanding *skill* to our content understanding agent to be able to keep up with the ever evolving policy landscape in a dynamic way.

Moving beyond basic pixel-level analysis, the agent evaluates multimodal semantic embeddings to isolate **"Generative Artifacts"**—cross-channel signatures of synthetic production. It targets repetitive "slop scripts" and structural visual inconsistencies indicative of automated generation. Furthermore, few-shot learning is integrated to identify emerging threats that violate the underlying intent of policies. The agent delivers a forensic classification (Authentic vs. Synthetic), identifies the specific modality of generative abuse (e.g., Synthetic Impersonation), and surfaces critical policy gaps for further review.

C. Behavior Understanding Agent (Inorganic Pattern Recognition)

The Behavior Understanding Agent is designed to interrogate the spatiotemporal footprint of suspected clusters to uncover hidden coordination markers. It scrutinizes infrastructure signals (such as ASN and device fingerprints) alongside inorganic activity patterns (including synchronized upload timestamps and "burst" publishing behaviors) that deviate significantly from organic human behavior. The output yields a comprehensive forensic summary detailing "inorganic" signatures (e.g., "100% of channels utilize identical OS versions and upload within the same 5-second window") to validate coordination.

D. Channel Cluster Understanding Agent

The Channel Cluster Understanding Agent is tasked with interrogating the structural connectivity and relationships within

the network. By analyzing inter-account linkages, it maps the full extent of coordinated "bot-nets," ensuring forensic investigations encompass the entire adversarial operation rather than isolated nodes.

The agent leverages a graph-based Relationship Signal service to surface established connections between channels within a suspect cluster. Furthermore, it conducts deep forensics on raw relationship signals to identify latent coordination markers among the channels. The output is a comprehensive forensic summary of detected inter-channel connections and infrastructure commonalities.

IV. Evaluation and Impact

We will use the following metrics to evaluate the performance of SAFE:

- **Accuracy**: Measured as agreement between the verdict generated by the root agent about a given set of channel IDs with the human analyst's verdict.
- **Increased Recall**: Measured as abusive trends that our system helps discover that previously would have gone unnoticed by existing ML classifiers and limited human reviewer capacity.
- **Efficiency**: Reduction of Average Handling Time (AHT) of investigating a cluster of channel IDs done by the human analyst.
- **Average Handling Time (AHT) Reduction**: In cases that the agent is not able to come up with a strong verdict, its insights can still expedite the human analyst investigations resulting in lower AHT.

V. Conclusion

The emergence of Generative AI has created a significant "synthetic gap," where traditional forensic workflows struggle to keep pace with the volume and velocity of automated abuse. SAFE addresses this challenge by providing a scalable multi-agent architecture that automates the forensic investigation process. By delegating specialized tasks to autonomous agents and synthesizing multimodal signals, SAFE significantly accelerates threat identification and enables the scaling of forensic operations to meet the demands of the modern adversarial landscape.

References

[1] D. Ghosh, W. Boettcher, R. Johnston, and S. Lahiri, "Bot Identification in Social Media," *TechRxiv*, DOI:10.36227/techrxiv.174062962.26956908/v1, 2025.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[3] P. Kirdemir, T. Adeliyi, and N. Agarwal, "Detecting and Characterizing Inorganic User Engagement on YouTube," in *Proceedings of the 18th International Conference on Web and Social Media (ICWSM) Workshops*, 2024.

[4] D. Schwarz, "DeepAgent: A Dual Stream Multi Agent Fusion for Robust Multimodal Deepfake Detection," *arXiv:2503.23642 [cs.CV]*, 2025.

[5] J. Ma et al., "Multimodal Transformers for Nuanced Policy Violation Detection in Digital Media," *Journal of AI Safety & Security*, 2024.

[6] L. Zhao and A. Gupta, "Semantic Safety Nets: Real-time Policy Enforcement using Retrieval-Augmented Transformers," in *Proceedings of the International Conference on Computational Forensics*, 2025.

[7] J. Yu, X. Xie, Q. Hu, Y. Ma, and Z. Zhao, "Chimera: Harnessing Multi-Agent LLMs for Automatic Insider Threat Simulation," in *Proceedings of the 32nd Network and Distributed System Security (NDSS) Symposium*, 2025.