Detection and Measurement of Near-Policy-Violating Content in Online Platforms

Armando Ordorica
Pinterest
New York, NY, USA
aordorica@pinterest.com

Anna Kiyantseva
Pinterest
San Francisco, CA, USA
akiyantseva@pinterest.com

Karim Wahba
Pinterest
San Francisco, CA, USA
kwahba@pinterest.com

Yucheng Tu
Pinterest
Seattle, WA, USA
ytu@pinterest.com

Adam Avery
Pinterest
Phoenix, AZ, USA
aavery@pinterest.com

Abstract

Slop refers to low quality, borderline inappropriate content that skirts policy violations. While slop has long posed challenges for content platforms, the rise of generative AI threatens to increase its prevalence by making it easier to efficiently produce large volumes of visually compelling content that may be misleading or low in informational value. This trend threatens user trust and the long-term health of recommendation systems. In this paper, we present a principled and scalable framework for detecting and measuring slop in the absence of explicit policy labels. Our approach has three main components. First, we introduce a detection method that applies relaxed thresholds to content signals and expands coverage via embedding-based similarity, capturing visually and semantically related items that may evade direct detection via signals. Second, we introduce tools for measuring the distributional skew of slop across the platform, at both the user and image signature level, using Lorenz curves and Gini coefficients to quantify its concentration. Third, we develop measurement techniques for estimating user affinity to slop content using proxy engagement signals. This is particularly useful in sparse signal spaces, such as for low activity users or rare content types, where explicit feedback is limited. We introduce a TF-IDF-inspired persona score that leverages impression logs to estimate user level interest without requiring explicit actions like saves or clicks. Together, these methods offer a comprehensive framework for detecting borderline content and measuring its distribution and consumption patterns, laying the groundwork for future mitigation efforts in recommender systems.

Keywords

recommender systems, personalization, borderline content, proxy metrics, low-signal users

ACM Reference Format:
Armando Ordorica, Anna Kiyantseva, Karim Wahba, Yucheng Tu, and Adam Avery. 2025. Detection and Measurement of Near-Policy-Violating Content in Online Platforms. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Slop refers to low-quality, borderline-inappropriate, or disruptive visual content that slips past automated filters and degrades the user experience. It is not overtly policy-violating, but is typically visually or semantically jarring, engagement-hacking, or low-effort AI-generated spam [15]. **Throughout this paper, we use the terms *gray zone content*, *slop content*, and *borderline content* interchangeably to refer to material that is not explicitly policy-violating but can still be detrimental to user experience and platform health.** Although slop is not exclusive to GenAI, recent advances in generative AI have significantly exacerbated the proliferation of slop, amplifying risks to recommendation quality [22]. Slop threatens user trust, degrades ecosystem health, and introduces noise into measurement pipelines. This phenomenon has drawn significant media attention, with reports highlighting the surge of AI-generated spam and its impact on social media platforms and the broader web [4, 31].

Slop can manifest at different levels:

- **Image level:** suggestive, bizarre, or uncanny imagery; click-bait visuals; algorithmically manipulated compositions designed to attract attention [7, 24].
- **Creator level:** accounts that mass-produce low-effort, AI-generated, or borderline content for engagement farming [16, 29, 32].
- **Metadata level:** misleading thumbnails, deceptive captions, keyword stuffing, or links to low-quality or irrelevant landing pages [3, 27, 38].

Examples of what could be considered slop are shown in Table 1.

In this work, we propose a principled framework for detecting and measuring slop by leveraging a combination of content signals, embedding-based proximity, and engagement data. Our approach is specifically designed for use in settings where explicit policy definitions of such content are lacking or evolving, making it practical for companies or teams facing ambiguous moderation boundaries. By leveraging embeddings, our framework supports ongoing measurement and enables a dynamic, data-driven definition of "slop" that

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Clearly Slop	Unclear	Clearly Not Slop
Racy content	Memes	Inspiring pictures
AI-generated content with misleading links		High-reputation articles
Medical misinformation		

Table 1: Examples of content types across the slop spectrum.

can flexibly expand as new forms of content emerge. Additionally, we introduce a method to infer user affinity to slop in the absence of explicit negative feedback, enabling more scalable and personalized mitigation strategies.

2 Literature Review

Work across academia and industry has long highlighted the growing concern around *gray zone content*: material that skates the edge of platform policy, often engineered to maximize engagement while avoiding overt violations [10, 19]. Platforms such as Facebook and YouTube have acknowledged the prevalence and risks of such content in public statements, noting that engagement tends to spike as content approaches policy thresholds and have responded with “reduce, not remove” interventions (e.g., downranking engagement bait) rather than solely relying on binary removal [11, 17, 35]. Recent detection techniques increasingly leverage embedding-based and multi-modal models, combining textual, visual, and graph signals to identify subtle and low-effort adversarial content [6, 37]. Frameworks that integrate user feedback and engagement signals to predict regret or negative sentiment have also been proposed [18, 36]. Yet, the rapid proliferation of generative AI has introduced new forms of low-effort, visually or semantically jarring 'slop' content that slips past traditional filters and defies clear categorization. Our work addresses these challenges by proposing a principled framework for detecting and characterizing such content using internal content signals, embedding proximity, and behavioral proxies for user affinity.

3 Defining Slop

In the absence of explicit ground-truth labels for slop, we construct proxy labels using a curated set $\mathcal{S} = S_1(x), S_2(x), \ldots, S_k(x)$ of k content signals, where each S_j is applied to a content item x to flag potentially low-quality or borderline content, with j indexing the k signals ($j = 1, \ldots, k$). These signals include detectors for gross or gruesome imagery, racy content, and other categories strongly correlated with negative user feedback (e.g., hides, reports).

For each signal, we select a detection threshold τ_j, forming a set of thresholds $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_k\}$, where $|\mathcal{T}| = |\mathcal{S}| = k$. Each threshold τ_j is determined by analyzing the empirical probability density function (PDF), $f_j^{\log}(\theta)$, of the log-transformed signal values $\log S_j(x)$ over all content items and locating modes associated with slop content. Because these empirical distributions are often multimodal in practice [2], we focus on the mode $\mu_j^{(\text{borderline})}$ corresponding to the borderline content cluster. The threshold τ_j for each signal S_j is then defined by

$$\tau_j = \arg\max_{\theta<\mu_j^{(\text{borderline})}} f_j^{\log}(\theta), \tag{1}$$

where $f_j^{\log}(\theta)$ is the estimated PDF of $\log S_j(x)$, and $\mu_j^{(\text{borderline})}$ denotes the location of the borderline mode. This empirical, mode-based thresholding ensures that relaxed thresholds capture a broader, but still relevant, set of candidates for slop detection. Note that these relaxed thresholds are intentionally more inclusive than those used for strict content filtration, allowing for the identification of items that may not warrant removal but could still degrade the user experience.

After finding these thresholds, each **signal is binarized** as:

$$B_j(x_i) = \mathbb{1}[S_j(x_i) > \tau_j], \tag{2}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function.

We define slop content in two complementary ways:

- **(1) Rule-based union:** Any item is slop if at least one (binarized) signal is triggered:

$$\text{Slop}_{\text{union}}(x) = \max_{j=1}^{k} B_j(x) = 1. \tag{3}$$

- **(2) Slop "core" (maximal intersection):** We define the total count of triggered signals for x as $C(x) = \sum_{j=1}^{k} B_j(x)$, and define the “core” set as:

$$\text{Core} = \left\{ x \in X \mid C(x) = \max_{x' \in X} C(x') \right\}. \tag{4}$$

We then compute the centroid of these core items within a Pin-level embedding space. Specifically, each item x is mapped to an embedding vector $\phi(x) \in \mathbb{R}^d$, where ϕ is the Pin-level embedding encoder. The centroid μ is given by:

$$\mu = \frac{1}{|\text{Core}|} \sum_{x \in \text{Core}} \phi(x). \tag{5}$$

The Pin-level embedding integrates both node connectivity and content-based features, allowing us to generate semantically meaningful and scalable content representations [21]. We subsequently classify as slop any item y located within a radius α of this centroid in the embedding space:

$$\|\phi(y) - \mu\| \leq \alpha \tag{6}$$

using Facebook AI Similarity Search (FAISS) [23] for efficient nearest-neighbor retrieval at scale.

Future iterations may incorporate richer representations such as GPT-based text embeddings, visual embeddings, or multimodal models like CLIP. While visual LLMs can help detect slop at the image level, they are limited in capturing higher-order context such as creator intent and domain reputation, as emphasized by Meta in their recent work on the safety of generative models [1]. The rule-based method is transparent and easily interpretable but may miss subtle cases. The embedding-based approach captures more nuanced, semantically similar content but may surface less interpretable edge cases, requiring careful tuning of the radius parameter.

We also incorporate a $(k + 1)^{\text{th}}$ signal, $S_{k+1}(x)$, generated by an in-house classifier designed to detect content likely created by generative AI (GenAI). However, given that GenAI content can also deliver genuine user value (e.g., ideas for home decor, hairstyles, or nail art), we construct two operational variants of slop:

$$\text{Slop}^{+}(x) = \max\left(\max_{j=1,\ldots,k} B_j(x),\ B_{k+1}(x)\right), \tag{7}$$

$$\text{Slop}^{-}(x) = \max_{j=1,\ldots,k} B_j(x), \tag{8}$$

where $B_j(x)$ is the binarized output for each content signal, as previously defined. Here, $\text{Slop}^{+}(x)$ defines the upper bound, in which GenAI content is classified as slop if detected, while $\text{Slop}^{-}(x)$ defines the lower bound, ignoring the GenAI signal. This dual-definition framework characterizes the slop region with upper and lower bounds, reducing the risk of overfitting to a narrow or overly deterministic boundary. The goal of this initial definition of slop (v0) is to establish a *functional* definition, not a perfect policy, to enable measurement of prevalence and consumption of this content type.

4 Measuring Concentration of Slop Content

To accurately analyze and visualize slop content, it is important to distinguish between two key units of measurement within our platform: the *Pin id* and the *image signature.* A Pin id uniquely identifies a specific instance of content as it appears on the platform. Each time a user uploads or saves a piece of content, it is assigned a distinct Pin id, even if the image itself is visually identical to others. Thus, two users saving the same photo will yield two different Pin ids. In contrast, an image signature is a hash or embedding that captures the underlying visual content, irrespective of duplication or repins. This signature-based approach is standard in industry, used by companies like Meta [28] and Google [8] for deduplication, efficient image storage, and image-based retrieval.

Building on observations from prior studies, which found that low-quality or low-credibility content can achieve disproportionate reach or engagement [9, 13, 14], we hypothesized that slop content would be more concentrated than non-slop content. In other words, we expected that a small subset of items or users would account for a large share of slop exposure. Understanding the degree of this concentration can inform both the propagation dynamics of slop and the design of potential interventions: if the distribution is highly concentrated, targeted solutions may be effective, whereas more even distributions could require broader, systemic strategies. **Prior work has used similar distributional analysis—measuring how content or engagement is concentrated among a small subset of users or items—to study virality and influence in online social networks** [34]. Below we introduce two key statistical tools to quantify and visualize this concentration: Lorenz curves [25] and Gini coefficients [20].

Lorenz Curves. A Lorenz curve is a graphical tool used to visualize the degree of inequality within a distribution. Originally developed in economics to represent the concentration of wealth or income [25], Lorenz curves have since been widely adopted to study disparities in many domains, including social media engagement [5, 34]. The curve plots the cumulative proportion of a quantity (such as slop impressions) against the cumulative proportion of the population (such as users or image signatures), ordered from lowest to highest contributor. If every entity contributed equally, the Lorenz curve would be a straight diagonal line; the more the curve bows away from this line, the greater the inequality or skew in the distribution [34]. Lorenz curves help reveal whether slop is highly concentrated within a small group, which are critical insights for designing effective interventions.

Gini Coefficients. Originally developed to quantify income and wealth inequality in economics [20], the Gini coefficient serves as a single-number summary of how concentrated slop exposures are across users or images. A higher Gini coefficient implies that a small minority accounts for a large share of slop impressions, signaling high concentration, while a lower value indicates a more even spread across the population [5, 34]. This allows for easy comparison between different types of slop signals and provides actionable insight into whether interventions should target a narrow set of contributors or be applied more broadly [5, 30].

Analysis using Lorenz curves (Figure 1) and the Gini coefficient (Figure 2) shows that slop content is indeed more concentrated than non-slop content, with a relatively small subset of image signatures accounting for a disproportionate share of slop impressions. However, the degree of concentration is not so extreme that targeting only a handful of image signatures would be an effective intervention strategy; a nontrivial number of distinct image signatures would still need to be addressed.

Figure 3 presents a top-K analysis, showing the cumulative share of total impressions and repins attributable to the top $x\%$ of users. Two notable patterns emerge. First, a small fraction of users accounts for a large share of activity for both metrics, but this effect is even more pronounced for repins: the top $x\%$ of users drive a larger percentage of repins than impressions. Second, this concentration is greater for slop content compared to non-slop content. Specifically, for any given top $x\%$ of users, a larger proportion of slop impressions and slop repins can be attributed to them versus non-slop. This indicates that not only are repins more concentrated than impressions, but engagement with slop content is even more heavily skewed among a small subset of users.

5 Characterization of Slop Affinity Using Implicit Engagement Signals

Slop, in its definition of non-policy violating content, is not a filtering problem but a personalization one. This perspective aligns with approaches taken by platforms like Meta, YouTube, and TikTok, which emphasize personalization approaches over hard filtering. These systems aim to calibrate who sees such content and when, rather than removing it outright [17, 35]. However, determining which users we should show slop content to vs not is nontrivial. The notion of generic "slop enjoyer" vs "slop non-enjoyers" is reductive. Individual users can fluctuate in their affinity for "slop" content, not just across topics, but even within the same topic across multiple sessions. Thus, characterizing the engagement patterns and the types of content that lead to "enjoyment" (or lackthereof) calls for a more nuanced and systematic approach. Our findings indicate that segmenting users along a continuous spectrum of engagement scores yields better results than applying interventions based on coarse, discrete user or content groupings.

Ideally, explicit signals like repins can help us determine whether a user “welcomes” a particular type of slop. However, few users

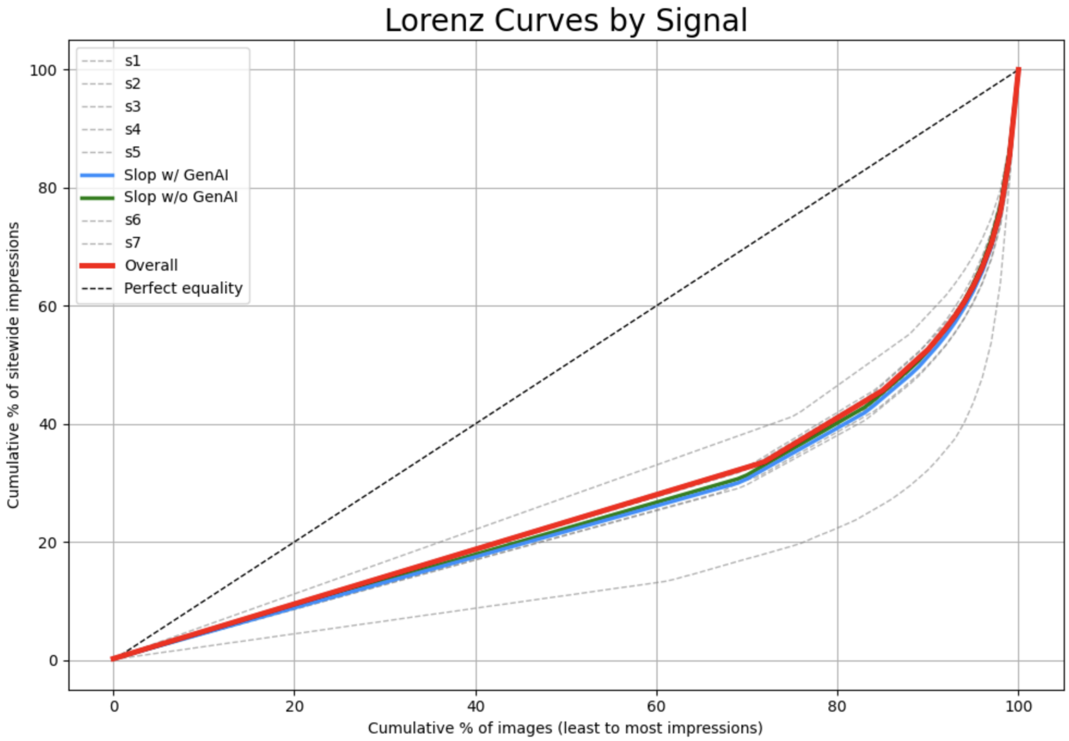

Figure 1: Lorenz curves for slop related signals. Each curve shows the cumulative share of impressions received by the bottom $x\%$ of image signatures, quantifying the inequality of different slop-related signals. The "Overall," "Slop w/ GenAI," and "Slop w/o GenAI" curves are highlighted for reference.

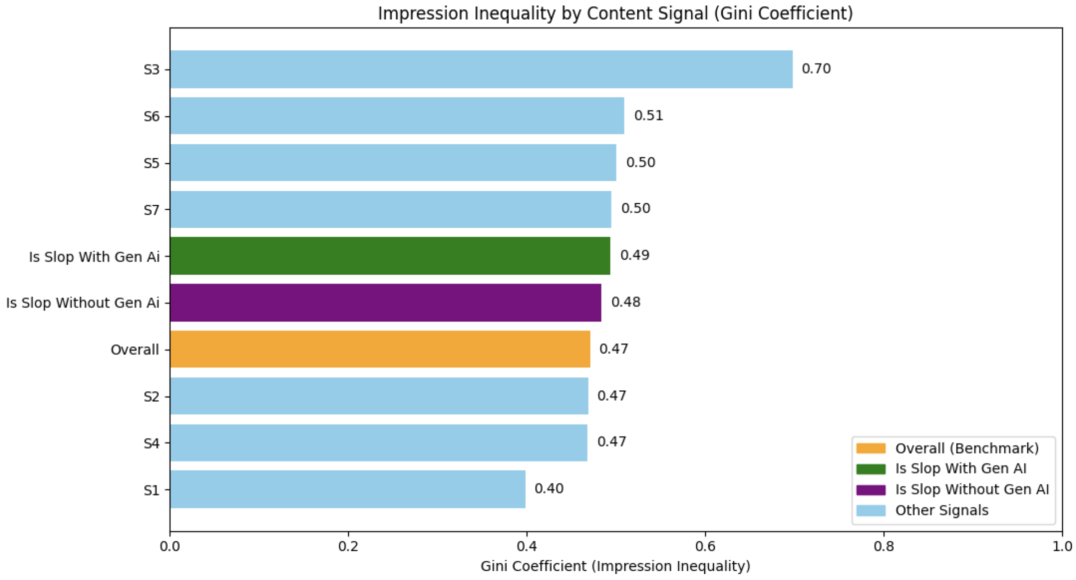

Figure 2: Gini coefficients for obfuscated slop-related signals. Each bar reflects the degree of concentration for slop exposure associated with a given signal: higher Gini values indicate that slop impressions are more heavily concentrated among a small subset of image signatures. The "Overall," "Slop w/ GenAI," and "Slop w/o GenAI" bars are highlighted for reference.

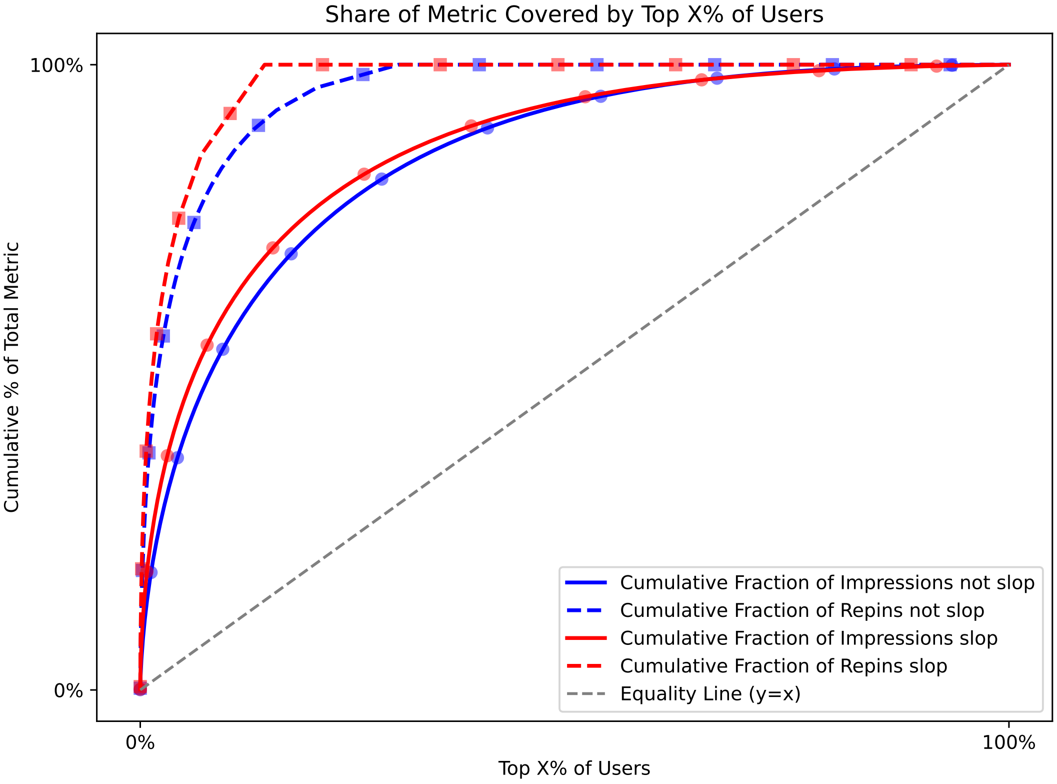

Figure 3: Top K Analysis. This shows the share of total impressions and repins attributable to the top $x\%$ of users. Both metrics are highly concentrated, but the skew is stronger for repins than impressions, indicating that a disproportionately small set of users drives the majority of repin activity on the platform.

save Pins on any given day, and even fewer actively hide or report content. Because of this scarcity of explicit feedback signals, proxy methods that leverage implicit signals, such as scrolling behavior and impressions to estimate user affinity to slop becomes essential, especially for low-signal users.

To address sparse interaction data, we propose a transformation inspired by Natural Language Processing (NLP), specifically, Term Frequency-Inverse Document Frequency (TF-IDF). In NLP, TF-IDF is a lightweight feature extraction technique that identifies terms that are unusually prominent within a document relative to their frequency across the corpus [26, 33]. This allows it to isolate the most distinctive signals while downweighting common, less informative terms, effectively denoising the representation [26]. The underlying intuition is rooted in information theory. The more surprising or rare a term is in the global context, the more informational value it carries when it appears, making it especially useful for characterizing the unique content of a document [12]. Inspired by these principles, we developed an approach to estimate a user's affinity to slop content by leveraging their impression history. We can treat user sessions as documents and content categories (e.g., slop-related clusters or topics) as terms. TF-IDF then allows us to highlight content categories that are shown more frequently in a given user's sessions than they are across the general population. This results in a **persona affinity score** that captures how **uniquely and intensely a user is exposed to slop content**, normalized by their overall exposure. While passive behavior patterns may be subtle, the distribution and concentration of impressions across content categories encode rich behavioral signals that relying exclusively on direct engagement can overlook. **Crucially, this measure does not require explicit engagement feedback, making it especially useful for low-signal users who may not have repins or clicks.** When benchmarked against affinity scores derived from explicit interactions, such as repins, **the TF-IDF-based scores demonstrate approximately 90% overlap, suggesting strong alignment and validating its use as a scalable proxy for slop affinity.**

While it is reasonable to be concerned that impression-based engagement might reinforce model biases, this worry should be alleviated by the fact that, for low-signal users, impressions on our platform are largely shaped by their explicit actions on Search and Related Pins. On these surfaces, what users see is determined more by their own intent than by what our recommendation systems predict they might want to see. On average, more than 70% of impressions originate from these two surfaces, both of which are directly triggered by user input [1]: in Search, impressions follow

[1] Pinterest Internal Data, accessed May 2025, users worldwide.

a text query, and in Related Pins, content is surfaced in response to closeups, where the viewed Pin acts as a query for visually or semantically similar content. Internal research further shows that users with infrequent Pinterest activity, such as those who visit episodically and often for specific, time limited purposes, rely on these interactive surfaces even more than regular users. As a result, for users who rarely repin or provide other forms of explicit feedback, TF-IDF-like transformations on impression data serves as a meaningful proxy for their interests, since it reflects their active navigation and intent rather than passive exposure dictated by the recommendation system.

5.1 Variant 1: Persona Affinity Score (without Global Prevalence Adjustment)

To quantify a user's affinity to slop content, we required a metric that was **simple, sensitive, with high coverage, comparable, and interpretable** that could be computed at scale. As such, we introduce two variants of the **persona score**, which are TF-IDF-inspired metrics that capture user affinity to slop content, scaled by the user's total volume of interactions.

Formula:

$$\text{Persona Score} = \frac{\text{Persona Metric}}{\text{Total Metric}} \times \log(1 + \text{Total Metric}) \qquad (9)$$

Example: GenAI Pins. The **persona metric** is the proportion of GenAI Pins in a user session, calculated as

$$\text{Persona Metric}_{v1} = \frac{30}{100} = 0.30$$

and the **total metric** is the total number of impressions in the session, that is,

$$\text{Total Metric} = 100.$$

Interpretation: This formulation captures both the intensity of exposure, reflected in the proportion of slop content, and the volume of impressions, represented by the logarithm of the total number of impressions in a session. The use of $\log(1 + \text{Total Metric})$ ensures the score is always well defined and strictly positive, avoiding issues such as $\log(0)$, which is undefined. If the total metric is zero, meaning the session had no impressions, then $\log(1)$ evaluates to zero, and the overall persona score becomes zero as expected. This formulation also helps differentiate users who have the same proportion of slop exposure but very different impression counts. For example, a user who sees 3 out of 10 GenAI Pins is treated differently from a user who sees 3000 out of 10,000, even though both have the same proportion.

Score Comparison for 30% Racy Impressions:

- For 100 total impressions:

$$\text{Persona Score}_i = 0.3 \times \log(1 + 100) = 0.3 \times \log(101) \approx \mathbf{0.601}$$

- For 10,000 total impressions:

$$\text{Persona Score}_j = 0.3 \times \log(1 + 10000) = 0.3 \times \log(10001) \approx \mathbf{1.200}$$

Although both users have the same proportion of GenAI impressions (30 percent), the user with a larger total impression count receives a higher persona score. This reflects the intuition that exposure patterns based on larger sample sizes are more statistically reliable and less prone to noise. In other words, seeing 30 out of 100 impressions as GenAI may reflect chance or volatility, whereas 3,000 out of 10,000 is a more consistent and meaningful signal. The logarithmic scaling rewards larger volume without letting the score grow too aggressively, offering a balanced way to incorporate both strength and reliability of the signal.

Figures 4 and 5 illustrate how persona scores, calculated using either impressions or repins, enable a nuanced characterization of user affinity to slop content. Plot 1A in Figure 4 shows the distribution of persona scores across users when measured on impressions, revealing a highly skewed distribution in which most users cluster at low scores due to sporadic or absent GenAI exposure.

Plots 1B and 1C in Figure 4 further decompose this exposure by showing, respectively, the fraction of each user's impressions that are GenAI and the share of total GenAI impressions across the platform contributed by each decile. These plots demonstrate not only the prevalence but also the concentration of slop signals among higher-affinity users.

Complementary metrics constructed by applying the same transformation shown on Equation 9 to repins yields Figure 5. Here, we use repins as explicit signals of user interest, in contrast to impressions, which serve as implicit signals. Importantly, applying the same transformation enables meaningful comparison between these two types of engagement data: although repin-based distributions are coarser due to their greater sparsity, the overall user patterns remain consistent with those derived from impressions. This demonstrates that the transformation not only harmonizes explicit and implicit signals, but also preserves the key behavioral gradients necessary for robust user characterization.

5.2 Variant 2: Persona Affinity Score (with Global Prevalence Adjustment)

To account for the **global rarity of a signal**, we extend the basic persona score to include an inverse document frequency (IDF)-style scaling factor. This results in a formulation that combines **Affinity**, **Global Rarity**, and **Scale**.

Formula:

$$\text{Persona Score}_{v2} = \left(\frac{\text{Persona Metric}}{\text{Total Metric}} \times \text{IDF}\right) \times \log(1 + \text{Total Metric})$$

where

$$\text{IDF} = \log\left(\frac{\text{Total Users}}{\text{Users with Persona Impression}}\right)$$

Interpretation: This score emphasizes rare behaviors that are highly concentrated in certain users, by scaling the persona metric based on how uncommon the behavior is across the user base.

While variant 2 has not been significantly studied and experimented with, it is a promising direction for future work. It allows us to capture the **uniqueness of a user's engagement with slop content relative to the overall population**, which can be particularly useful for identifying **niche interests or emerging trends** that may not be widely recognized yet.

6 Conclusion and Future Work

We presented a scalable, principled framework for detecting and measuring "slop," defined as borderline, low-quality content that

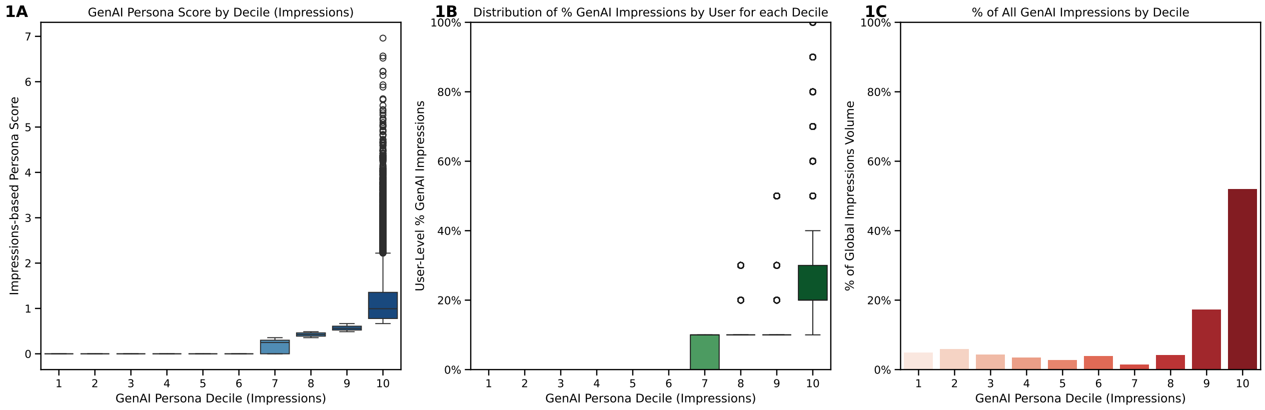

Figure 4: This figure shows the distribution of persona scores for users in each decile, using variant 1 of the persona score calculation and using impressions as units. The distribution is highly skewed: because most users do not have any GenAI impressions, the majority have a persona score of zero and are concentrated in the lower deciles (1 to 6), whereas only a small fraction of users (deciles 7 to 10) have substantially higher scores.

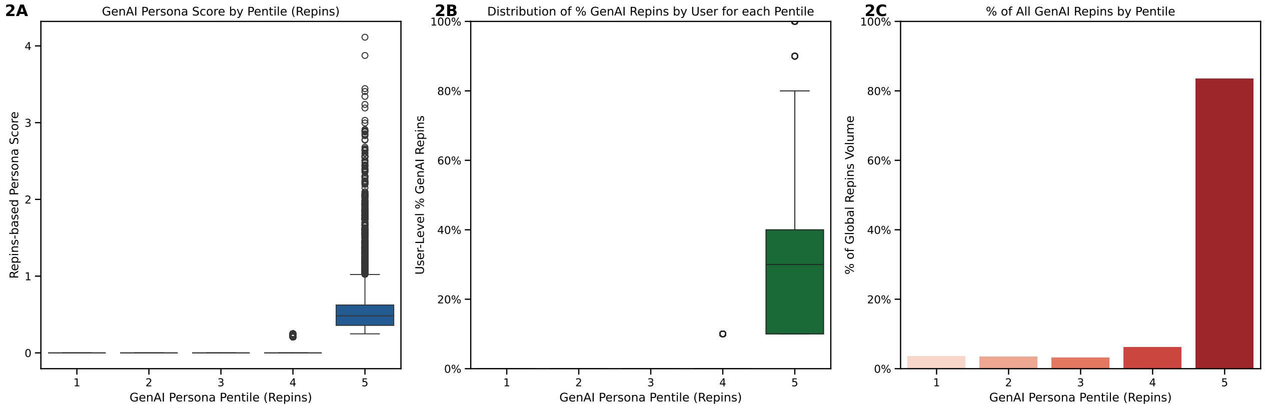

Figure 5: This figure shows the distribution of persona scores for users in each pentile, using variant 1 of the persona score calculation and repins as units. The distribution is highly skewed: because most users do not repin GenAI content, the majority of them have a persona score of zero and are concentrated in the lower pentiles (1 to 3), whereas only a small fraction of users (pentiles 4 and 5) have substantially higher scores.

evades policy violations but can undermine user experience and trust. Our method combines relaxed thresholds on content signals with embedding-based expansion and leverages TF-IDF-style transformations on impression-level logs to provide a robust, operational definition of slop. This approach enables accurate measurement of prevalence and user affinity, including for users with little explicit engagement.

Our experiments show that although slop is measurably more concentrated than non-slop content, the skew is not extreme enough to justify naive interventions targeting only a small set of items or users. Notably, we demonstrate that users missed by traditional engagement metrics can nonetheless be effectively profiled with scalable, TF-IDF-based persona scores built from impression histories.

Future work should investigate how slop content achieves virality and whether it exploits unique dynamics or "blind spots" in recommendation systems that are less available to high-quality content, including the role of feedback loops and ranking signals. Another important direction is to evaluate intervention strategies that personalize slop exposure at different stages of the recommendation stack—including retrieval, filtering, ranking, and post-ranking calibration—rather than focusing solely on demotion or filtering. Additionally, advancing detection through richer multimodal representations, studying long-term outcomes of slop mitigation, and incorporating signals such as creator reputation or cross-cohort engagement can further improve the robustness and fairness of future recommender systems.

References

[1] Meta AI. 2023. Towards Safer Generative Models. https://ai.meta.com/blog/towards-safer-generative-ai/. Accessed: 2025-05-26.

[2] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery* 29, 3 (2015), 626–688.

[3] Muaaz Al-Sarem, Faten Saeed, Khaled Alzahrani, Abdullah Alsabaani, Khaled Omar, and Seifedine Kadry. 2021. Ads and Fraud: A Comprehensive Survey of Fraud in Online Advertising. *Computers* 10, 4 (2021), 39. doi:10.3390/computers10040039

[4] Justin Bariso. 2023. *Social Media Is Being Flooded With Spammy AI Content.* Business Insider. https://www.businessinsider.com/social-media-flooded-spammy-ai-content-2023-8 Accessed: 2025-05-27.

[5] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11 (2016). https://firstmonday.org/ojs/index.php/fm/article/view/7090

[6] Akash Bonagiri, Lucen Li, Rajvardhan Oak, Zeerak Babar, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. Towards Safer Social Media Platforms: Scalable and Performant Few-Shot Harmful Content Moderation Using Large Language Models. arXiv:2501.13976 [cs.CL] https://arxiv.org/abs/2501.13976

[7] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. 2020. Amplifying The Uncanny. arXiv:2002.06890 [cs.CV]

[8] Bingyi Cao and Mário Lipovský. 2022. Introducing the Google Universal Image Embedding Challenge. https://research.google/blog/introducing-the-google-universal-image-embedding-challenge/ Accessed: 2025-05-26.

[9] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 9–16. doi:10.1109/ASONAM.2016.7752207

[10] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, Vol. 1. 31:1–31:22. doi:10.1145/3134666

[11] Josh Constine. 2018. Facebook will change algorithm to demote "borderline content" that almost violates policies. https://techcrunch.com/2018/11/15/facebook-borderline-content/. https://techcrunch.com/2018/11/15/facebook-borderline-content/ Accessed: 2025-05-26.

[12] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience.

[13] Qiang Cui, Sabrina Gaito, and Giancarlo Ruffo. 2022. Measuring user engagement with low credibility media sources in a controversial online debate. *EPJ Data Science* 11, 1 (2022), 45. https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-022-00342-w

[14] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour* 1, 7 (2017), 1–7. https://arxiv.org/abs/1701.02694

[15] Renée DiResta and Josh A. Goldstein. 2024. How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. *Harvard Kennedy School Misinformation Review* (2024). https://misinforeview.hks.harvard.edu/article/how-spammers-and-scammers-leverage-ai-generated-images-on-facebook-for-audience-growth/ Accessed: 2025-05-27.

[16] Renée DiResta and Josh A. Goldstein. 2024. How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. *CoRR* abs/2403.12838 (2024). arXiv:2403.12838 https://arxiv.org/abs/2403.12838

[17] Facebook. 2017. News Feed FYI: Fighting Engagement Bait on Facebook. https://about.fb.com/news/2017/12/news-feed-fyi-fighting-engagement-bait-on-facebook/ Accessed: 2025-05-26.

[18] Lluís Garcia-Pueyo, Vinodh Kumar Sunkara, Prathyusha Senthil Kumar, Mohit Diwan, Qian Ge, Behrang Javaherian, and Vasilis Verroios. 2023. Detecting and Limiting Negative User Experiences in Social Media Platforms. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery, 4086–4094. doi:10.1145/3543507.3583883

[19] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8, 3 (2022). https://www.researchgate.net/publication/362800371_Do_Not_Recommend_Reduction_as_a_Form_of_Content_Moderation Accessed: 2025-05-26.

[20] Corrado Gini. 1912. Variabilità e mutabilità. *Studi Economico-Giuridici della R. Università de Cagliari* 3 (1912), 3–159. Accessed: 2025-05-27.

[21] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, Vol. 30. https://arxiv.org/abs/1706.02216

[22] John Herrman. 2024. *The Internet's AI Slop Problem Is Only Going to Get Worse.* https://nymag.com/intelligencer/article/ai-generated-content-internet-online-slop-spam.html Accessed: 2025-05-27.

[23] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. https://github.com/facebookresearch/faiss. Accessed: 2025-05-26.

[24] Vivek Kaushal, Sawar Sagwal, and Kavita Vemuri. 2022. Clickbait's Impact on Visual Attention – An Eye Tracker Study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44. https://escholarship.org/uc/item/8w80h7jp

[25] Max O. Lorenz. 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association* 9, 70 (1905), 209–219. https://www.jstor.org/stable/2276207

[26] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press. https://nlp.stanford.edu/IR-book/

[27] Arunesh Mathur, Gunes Acar, Michael G. Friedman, Elena Lucherini, Jonathan R. Mayer, Marshini Chetty, Arvind Narayanan, and Steven Engelhardt. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, Vol. 3. 1–32. doi:10.1145/3359183

[28] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. 2022. A Self-Supervised Descriptor for Image Copy Detection. arXiv:2202.10261 [cs.CV] https://arxiv.org/abs/2202.10261

[29] Giovanni Puccetti et al. 2024. AI 'News' Content Farms Are Easy to Make and Hard to Detect: A Case Study in Italian. *arXiv preprint* (2024). arXiv:2406.12128 https://arxiv.org/abs/2406.12128

[30] Filippo Radicchi, Claudio Castellano, Fabio Cecconi, Vittorio Loreto, and Domenico Parisi. 2004. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences* 101, 9 (2004), 2658–2663. doi:10.1073/pnas.0400054101

[31] Adi Robertson. 2023. *AI is being used to generate whole spam sites.* The Verge. https://www.theverge.com/2023/5/2/23707788/ai-spam-content-farm-misinformation-reports-newsguard Accessed: 2025-05-27.

[32] Tate Ryan-Mosley. 2023. Next-gen content farms are using AI-generated text to spin up junk websites. *MIT Technology Review* (2023). https://www.technologyreview.com/2023/06/27/1075545/next-gen-content-farms-ai-generated-text-ads/

[33] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523. doi:10.1016/0306-4573(88)90021-0

[34] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality Prediction and Community Structure in Social Networks. *Scientific Reports* 3 (2013), 2522. doi:10.1038/srep02522

[35] YouTube Official Blog. 2019. The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/ Accessed: 2025-05-26.

[36] Chenghui Yu, Peiyi Li, Haoze Wu, Yiri Wen, Bingfeng Deng, and Hongyu Xiong. 2024. USM: Unbiased Survey Modeling for Limiting Negative User Experiences in Recommendation Systems. *arXiv preprint arXiv:2412.10674* (2024). https://arxiv.org/abs/2412.10674

[37] Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. 2025. ShieldGemma 2: Robust and Tractable Image Content Moderation. arXiv:2504.01081 [cs.CV] https://arxiv.org/abs/2504.01081

[38] Herbert Zuze and Melius Weideman. 2013. Keyword Stuffing and the Big Three Search Engines. *Online Information Review* 37, 2 (2013), 228–238. doi:10.1108/OIR-01-2012-0005