

Does AI Amplify Online Hate?

A Multi-Modal Measurement Study on 4chan's /pol/

Bo Shao
CISPA Helmholtz Center for
Information Security,
Saarbrücken, Germany
bo.shao@cispa.de

Xinyue Shen
University of Waterloo,
Waterloo, Canada
CISPA Helmholtz Center for
Information Security,
Saarbrücken, Germany
xinyue.shen@uwaterloo.ca

Yang Zhang*
CISPA Helmholtz Center for
Information Security,
Saarbrücken, Germany
yangzhang.almo@outlook.com

Abstract

As generative AI models gain widespread adoption, an urgent question arises: are they being exploited to produce hateful content in online communities? We present the first large-scale, multi-modal measurement study to investigate this question, drawing on 465M text posts spanning 2016-2026 and 16.9M images spanning 2021-2025 from 4chan's /pol/ board. Specifically, we focused on two hate targets with long-established phenomena in parallel: anti-Indian hate and anti-Semitic hate. By benchmarking five text and five image AI detectors on hateful content, we deploy the best-performing detectors on over 31K long-form hateful posts and approximately 50K hateful images to measure AI-generated content prevalence. We find that the two targets exhibit different growth patterns: anti-Indian hate is substantially self-sustaining, with 46.15% of detected peaks driven by organized community campaigns rather than external events, against only 5.88% for anti-Semitic hate, which is instead largely reactive, with 94.12% of peaks triggered by real-world events such as the October 7 Hamas attack. AI-generated hateful posts have grown steadily since late 2022, reaching 1.42% of long-form posts by 2025. More importantly, AI adoption diverges across the two targets. For anti-Indian hate, AI content prevalence correlates significantly with slur frequency, nearly doubles during hate peaks, and AI-generated posts are 55.72% more toxic than human-written ones. Anti-Semitic hate, by contrast, shows no significant AI-hate correlation, indicating that AI tools are more widely adopted in campaign-driven hate than in established hate discourse. Lead-lag analysis reveals that hate campaigns precede AI adoption by approximately three months, suggesting that generative AI serves as an accelerator for existing hate movements rather than an independent driver.

Disclaimer. This paper contains examples of hateful content. Reader discretion is recommended.

CCS Concepts

• Security and privacy → Social aspects of security and privacy.

*Yang Zhang is the corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. CCS '26, The Hague, Netherlands

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2871-6/2026/11

<https://doi.org/10.1145/3830454.3846752>

Keywords

AI-Generated Content, Hate Speech, Online Extremism

ACM Reference Format:

Bo Shao, Xinyue Shen, and Yang Zhang. 2026. Does AI Amplify Online Hate? A Multi-Modal Measurement Study on 4chan's /pol/. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26)*, November 15–19, 2026, The Hague, Netherlands. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3830454.3846752>

1 Introduction

Large language models and text-to-image generators such as ChatGPT [27], Llama [49], and Stable Diffusion [39] have dramatically lowered the barrier to content production. However, these capabilities have also enabled the proliferation of hateful content production at an unprecedented scale. Recent work has shown that mainstream LLMs can be prompted with simple queries to generate hate speech targeting diverse identity groups [43]. Text-to-image generators have also been weaponized to produce hate imagery at scale in India [5] and anti-Semitic content targeting Jewish communities [2]. Compared to human-created hateful content, as illustrated in Figure 1, AI-generated hate speech is more fluent and incorporates a broad range of stereotypes, and AI-generated hate imagery exhibits higher resolution and intricate visual detail. This greatly amplifies the risks faced by online communities.

Prior work has primarily focused on measuring the prevalence of AI-generated content on social media [46] and developing detectors and benchmarks [10, 12, 24]. Regarding online hate, existing work has largely focused on measuring how real-world events, such as elections or communal violence, trigger spikes in hate speech and hateful memes targeting specific groups [30, 34, 42]. However, the impact of AI-generated content on hate spread remains poorly understood. This is a pressing concern for AI safety and content governance. Current models are deployed with safety mechanisms, including content filters and alignment training, designed to prevent harmful outputs. Controlled studies have shown that these mechanisms can be circumvented [3, 22, 43], and open-source model weights make it possible to remove safety alignment through local fine-tuning [33]. These studies establish that evasion is possible under laboratory conditions. Whether AI-generated hateful content actually appears at scale in online communities has not been empirically measured. This gap is especially pressing for 4chan's /pol/, an anonymous imageboard widely regarded as one of the most influential sources of online extremism [13, 30]. We therefore conduct a multi-modal measurement study of /pol/ to investigate



Figure 1: Four types of hateful content on /pol/: human-written text (a), AI-generated text (b), human-created images (c), and AI-generated images (d).

whether generative AI is amplifying online hate, including both *text* for its extremely low generation cost and *images* for their virality and ability to convey hateful stereotypes at a glance. We focus on anti-Semitic and anti-Indian hate because they are the two most discussed targets on /pol/. From 2022–2026, posts referencing each accounted for 4.27% and 1.70% of monthly posts, well above other frequently mentioned groups.

Specifically, we aim to answer three research questions:

- **RQ1:** How have hate speech and hate imagery targeting Indians and Jews on /pol/ changed over time, and what drives their temporal dynamics?
- **RQ2:** Has AI-generated content emerged in online hate, and how has its prevalence changed?
- **RQ3:** Is there evidence that AI-generated content co-occurs with and amplifies hate spread?

To answer **RQ1**, we track daily discussion volume and identify spikes in India-related and Jewish-related posts on /pol/, detecting 30 peaks across ten years (2016–2026). For each peak, we sample posts and extract top unigrams and bigrams to determine whether it was triggered by a real-world event or by an organized community campaign. We further discover slurs associated with each target group using Word2Vec embeddings trained on the full /pol/ corpus. To assess whether India- and Jewish-related posts are broadly more toxic beyond explicit slurs, we score posts using Google’s Perspective API [18]. For images, we build a semantic retrieval system based on CLIP [36] and FAISS [14] to query India-related and Jewish-related hate imagery from 16.9 million images and track their temporal trends. These RQ1 analyses operate on the full corpus of 465M text posts and 16.9M images to characterize hate discourse dynamics.

To answer **RQ2**, we measure the prevalence of AI-generated content in hateful posts and images. We benchmark five text detectors and five image detectors on domain-specific evaluations and select the best-performing detector for each modality: Pangram V3 for text (near-perfect precision of 0.995) and AIDE for images. We apply them to the hateful content retrieved from the full corpus, 31,756 long-form hate speech posts and 49,679 hate images to track prevalence over time.

To answer **RQ3**, we test whether AI-generated content and hate intensity are linked through four progressively stronger analyses: Spearman correlation between AI rates and slur frequency, comparison of AI rates during hate peak vs. non-peak months, toxicity comparison between AI-generated and human-written posts, and lead-lag cross-correlation analysis to determine the temporal direction of the relationship. We compare anti-Indian with anti-Semitic hate across all analyses to assess generalizability.

Main Findings. Our analysis yields the following main findings:

- (1) **Two hate targets, two growth patterns.** Anti-Indian hate is substantially self-sustaining: 46.15% of its peaks are driven by organized community campaigns such as cypasta [50], its slur vocabulary is expanding rapidly with “pajeeet” frequency increasing 7-fold from 2020 to 2025, and hateful imagery tripled from 0.07% to 0.23%. Anti-Semitic hate is reactive: 94.12% of its peaks are triggered by real-world events such as the October 7 Hamas attack, while its slur vocabulary remains stable (Section 4).
- (2) **AI-generated hateful posts have emerged and are growing.** Before ChatGPT’s release in November 2022, our detector flagged virtually no AI-generated posts, establishing a minimal false-positive baseline. Since then, AI-generated text has grown steadily, reaching 1.42% by 2025. AI-generated images trend slightly upward for anti-Semitic imagery (9.19% to 12.65%) but remain flat for anti-Indian imagery (Section 5).
- (3) **AI amplifies campaign-driven hate but does not initiate it.** For anti-Indian hate, AI-generated content correlates with hate intensity ($\rho=0.57$, $p<0.0001$), nearly doubles during hate peaks (1.39% vs. 0.78%), and AI-generated posts are 55.72% more toxic than human-written ones. Anti-Semitic hate shows no such correlation ($\rho=0.21$, $p=0.15$), indicating that AI tools are selectively adopted for campaign-driven hate. Lead-lag analysis shows that hate surges precede AI adoption by approximately three months, suggesting that generative AI serves as an accelerator for existing hate movements rather than a standalone motivator (Section 6).

2 Related Work

AI-Generated Text Detection. The rapid deployment of LLMs has spurred a growing body of work on machine-generated text detection. Early statistical approaches include Log-Likelihood thresholding, introduced by Solaiman et al. [44] alongside the release of GPT-2, and token-level Entropy analysis, proposed by Gehrman et al. [8] in the GLTR framework, both of which exploit distributional differences between human and machine text. DetectGPT [24] introduced a zero-shot method based on probability curvature. MGT-Bench [12] provides a systematic benchmark across multiple generators and detectors. Hans et al. [10] propose a zero-shot approach using perplexity ratios. Commercial services such as GPTZero [48] and Pangram [7, 29] provide proprietary detection solutions. Most prior evaluations focus on curated benchmarks, whereas our work is the first to apply AIGC detectors to hateful content in online communities.

AI-Generated Image Detection. Early work such as CNNDetection [53] has shown that CNN-generated images carry detectable artifacts. FerretNet [21] exploits local pixel dependencies for efficient

synthetic image detection, while C2P-CLIP [47] injects category-common prompts into CLIP to improve generalization in deepfake detection. More recent approaches like AIDE [77] combine high-level semantic and low-level artifact features for better generalization, while SAFE [19] applies simple image transformations to improve robustness. Ha et al. [9] evaluated both automated detectors and human judges on distinguishing human art from AI-generated images, finding that combining the two yields the best accuracy.

Hate Speech on 4chan's /pol/. 4chan's /pol/ board has been extensively studied as a nexus of online extremism. Hine et al. [13] provided the first large-scale measurement study of /pol/, characterizing its globally distributed user base and the prevalence of hate speech. Papasavva et al. [30] released a multi-year dataset of /pol/ posts. Zannettou et al. [79] traced the flow of alternative news from 4chan, finding that its alt-right communities exert surprising influence in spreading such content to mainstream platforms.

AI-Generated Hateful Content. A growing body of work examines how generative AI can produce hateful content. On the text side, Hartvigsen et al. [11] used GPT-3 to construct ToxiGen, a large-scale dataset of implicitly toxic statements targeting 13 minority groups, finding that annotators struggle to distinguish AI-generated text from human-written text. Shen et al. [43] demonstrated that LLM-generated hate speech systematically evades existing detectors. Model safety alignment can be circumvented through black-box jailbreaking [22] and fine-tuning that removes safety constraints [33]. On the image side, Qu et al. [35] found that text-to-image models generate unsafe images at rates up to 19%, and that fine-tuning via DreamBooth can produce hateful meme variants comparable to real-world examples. Ba et al. [3] achieved 88% success bypassing safety filters in commercial models like Midjourney. In the wild, recent reports document AI being used to generate anti-Indian and anti-Semitic content online [5, 40]. To the best of our knowledge, however, no prior work has systematically measured the prevalence and dynamics of AI-generated hateful content in online communities.

3 Data and Methodology

In this section, we first describe our data collection process, then present our hate measurement pipeline, including hate speech analysis and hate imagery analysis. Finally, we describe our approach to detecting AI-generated text and images. Figure 2 provides an overview of our data collection and analysis pipeline.

3.1 Data Collection

4chan is an anonymous image board organized by sub-communities called “boards,” each driven by a specific topic of interest. We focus on the Politically Incorrect board (/pol/) as it is the primary venue for political discussion and is known for spreading toxic and racist ideology [13, 30]. This has made 4chan's /pol/ a major platform for researchers to study online hate, extremism, and radicalization [13, 30, 79].

We obtain our data from 4plebs,¹ an unofficial archive of select 4chan boards (including /pol/) that preserves threads and

images in real time using the Asagi fetcher software. 4plebs publishes complete database dumps via the Internet Archive,² which we download and import into a PostgreSQL database for efficient querying. In total, our text dataset comprises 465,466,735 posts from /pol/ spanning 10 years (January 1, 2016 – January 10, 2026), enabling large-scale longitudinal analysis. For images, we download files referenced in posts from 4plebs' image archives. The image archive of /pol/ board alone, covering posts since September 2013, amounts to approximately 12 TB. We prioritize a subset of the data: our image dataset contains 16,945,311 images collected from March 29, 2021 to October 31, 2025. This range ensures our analysis covers the period both before and after AI image generation tools became widely available.

3.2 Hate Speech Analysis

To characterize the landscape of anti-Indian and anti-Semitic hate on /pol/, following previous research [42], we employ three complementary approaches. We first investigate the temporal dynamics of neutral reference terms (“india”/“indian” and “jews”/“jewish”) to map how discussion volume evolves and identify peak events. We then discover slurs via Word2Vec embeddings, revealing the semantic associations between each target and hateful language. Finally, we score post toxicity to assess severity beyond explicit slurs.

Trend Analysis. We track how often each target group is discussed on /pol/ by computing the proportion of posts containing neutral reference terms (“india”/“indian” and “jews”/“jewish”) for each day. Following previous research [42], to identify salient bursts, we apply peak detection using the `find_peaks` function from SciPy [52] on the daily proportions with a prominence threshold of 0.01 and a minimum distance of 14 days. For each detected peak, we estimate its temporal extent using full width at half maximum. To interpret these bursts, we employ *post-level content analysis*: for each peak, we sample 500 posts from the peak date ± 2 days and automatically extract the top-30 unigrams and top-15 bigrams by frequency (excluding stopwords and generic slurs). We use these keywords together with the peak date to trace the likely trigger of each peak. When keywords and timing jointly suggest a real-world event (e.g., “atul_subash” and “elon_musk” appearing at the peak on 2024-12-27, coinciding with the H-1B visa debate), we search news outlets and Wikipedia to verify the event, and label the peak as *event-driven* [E]. When keywords are instead dominated by platform-native content, i.e., organized campaigns and *copypasta* [50] (blocks of text that users repeatedly copy and repost verbatim), such as “fecalwall,” “codex_pajeet,” and “goyimtv,” and no corresponding external trigger can be identified, we label the peak as *discourse-driven* [D]. This event-vs-discourse classification was exploratory, added after observing the peak patterns. A second author independently verified all classifications.

Slur Discovery. While trend analysis captures discussion volume, it does not reveal the vocabulary of hate. To semi-automatically discover slurs towards a specific target group, we train Word2Vec models [23] on the full /pol/ corpus. Before training, we preprocess all posts by converting to lowercase, expanding contractions, and removing HTML tags, numbers, punctuation, and stopwords. We

¹<https://4plebs.org/>.

²<https://archive.org/details/@fourplebs>.

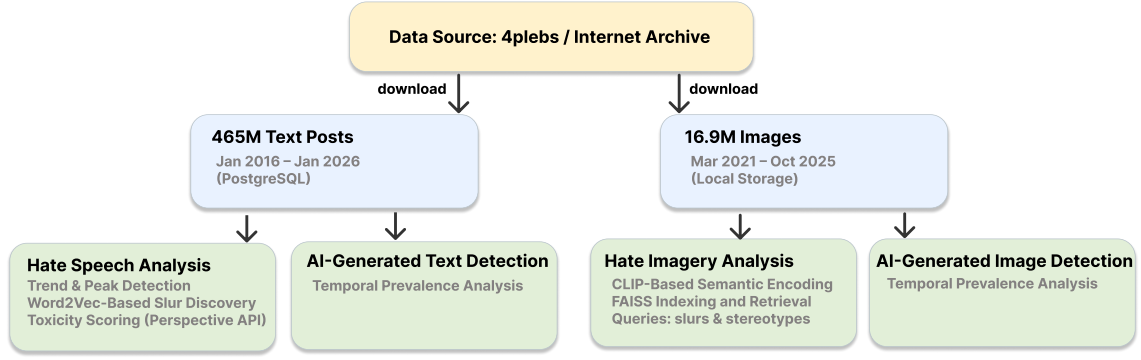


Figure 2: Overview of our data collection and analysis pipeline. Text posts are studied through trend analysis, Word2Vec-based slur discovery, and toxicity scoring. Images are encoded with CLIP for semantic retrieval of hate imagery. Both text and images are examined for AI-generated content using specialized detection methods.

Table 1: Top 15 most similar words to target terms, and top 15 most frequently occurring words towards combined hateful weight vectors. Values denote cosine similarity. Hateful terms are in bold.

Anti-Indian					Anti-Semitic				
india		indian		india/indian+hate	jews		jewish		jews/jewish+hate
inida	0.737	indain	0.660	inida	0.674	jewsih	0.696	kike	0.878
bangladesh	0.729	nepali	0.650	pooindoo	0.663	zionists	0.656	jewis	0.669
pakistan	0.703	indiam	0.649	shitskindia	0.653	hews	0.630	hewish	0.627
pajeetland	0.686	indan	0.649	pakistan	0.617	godkillers	0.605	jeiwsh	0.626
bangladesh	0.675	bangadeshi	0.644	currynigger	0.650	jwes	0.591	jewsish	0.626
china	0.674	inidan	0.639	bangladesh	0.633	muslims	0.588	merchantish	0.597
poojeetistan	0.665	bengali	0.637	idnian	0.602	joos	0.584	jewisj	0.578
pooindia	0.654	pakistani	0.628	dothead	0.673	arabs	0.583	jwish	0.575
pajeetistan	0.653	tamil	0.627	china	0.641	yids	0.575	jewidh	0.574
pajeetstan	0.653	imdian	0.621	indain	0.631	boshlevic	0.573	ewish	0.562
jeetland	0.647	punjabi	0.616	poojetistan	0.668	islamics	0.565	jew	0.555
pookistan	0.645	bangladeshi	0.616	streetshitter	0.600	muslimes	0.564	nonjewish	0.545
pakistand	0.639	pooindoo	0.605	indiam	0.592	sionists	0.560	jjewish	0.545
bengladesh	0.636	idnian	0.604	bangadeshi	0.579	jjews	0.560	jewisg	0.544
bangladesh	0.635	poojnabi	0.601	nepali	0.680	kykes	0.559	jawish	0.543
								nazi	0.842

train a Skip-gram model [23] (context window 5, minimum word count 20, 300 dimensions, 3 epochs) using Gensim [38] corpus file mode, yielding a vocabulary of 987K words.

Following prior work [13, 42], we leverage 15 widely-recognized hate words as seeds: *nigger*, *faggot*, *retard*, *bitch*, *idiot*, *cunt*, *kike*, *fag*, *nazi*, *trash*, *pussy*, *goy*, *frog*, *spic*, *chink*. For each seed-reference combination (e.g., “nigger” + “india”), we add their word embeddings and retrieve the 10 most similar words to the resulting vector. Iterating over all 15 seeds $\times$ 2 reference terms $\times$ 10 most similar words yields 300 candidate words per target group. We rank candidate words by the number of appearances and report the top 15 in Table 1. These slurs, together with established slurs documented in prior hate speech research [1, 6, 51], serve as the foundation for our downstream analyses.

Toxicity Analysis. The slur analysis captures explicit hate vocabulary, but hate speech also appears in posts that use neutral reference terms (“india”, “jews”) in derogatory or hostile ways without explicit slurs. To assess whether India- and Jewish-related posts are broadly more toxic, i.e., not just when slurs are present, we use Google’s

Perspective API [18] to score posts on two dimensions: SEVERE TOXICITY, which measures the degree of hate, rudeness, and disrespect, and IDENTITY ATTACK, which evaluates how negatively a post targets someone because of their identity (race, religion, nationality). Rather than applying a fixed classification threshold, which is difficult to justify given the subjectivity of toxicity, we report the cumulative distribution function (CDF) of scores to show the full distributional shift. We compare three mutually exclusive groups: India-related posts containing “india”/“indian”, Jewish-related posts containing “jews”/“jewish”, and a random baseline excluding all reference terms.

3.3 Hate Imagery Analysis

Since 4chan is fundamentally an imageboard, hateful content is not limited to text; memes, caricatures, and other visual media also play a central role in posts. To identify and track hate imagery at scale, we build a semantic retrieval system based on CLIP [36].

Embedding Computation. We encode all 16,945,311 images from /pol/ using the CLIP ViT-B/32 vision encoder [36]. Each image is

converted to RGB, resized and center-cropped to 224×224 pixels following CLIP’s standard preprocessing, and encoded into a 512-dimensional embedding vector. All vectors are L2-normalized so that inner products correspond to cosine similarity.

Index Construction. To enable efficient retrieval over 16.9M vectors, we build a FAISS [14] IVFPQ (Inverted File with Product Quantization) index, which supports approximate nearest-neighbor search with sub-second query latency.

Query and Retrieval. We design semantic search queries that capture both explicit slurs and broader hateful content. For anti-Indian content, we use three queries: “india” (neutral reference term), “pajeet” (primary slur from Table 1), and “anti indian meme” (explicit hateful content descriptor). For anti-Semitic content, we use: “jewish” (neutral reference term), “antisemitic caricature,” and “anti jewish meme.”

Given a text query, we encode it into the same 512-dimensional space using CLIP’s text encoder and retrieve all images whose cosine similarity exceeds a threshold of 0.27. We set this threshold by analyzing score distributions across queries and manually inspecting images near the boundary; below 0.27, retrieved images are predominantly unrelated to the query. This threshold-based approach allows each query to return a natural number of results (e.g., 19,540 for “pajeet” vs. 8,317 for “street shitter”), reflecting true differences in the volume of visual content. To account for fluctuations in overall posting activity, we normalize the weekly count of retrieved hate imagery by the total number of images posted on /pol/ that week, reporting rates as percentages rather than raw counts.

3.4 AI-Generated Text Detection

Existing AI text detectors are predominantly trained and evaluated on formal text genres such as news articles and academic essays. Their reliability on informal, adversarial, and slang-heavy text typical of 4chan’s /pol/ board remains unknown. To identify a detector suitable for the AI-generated hate text detection task, we construct a hate-speech-specific benchmark and compare five detection methods. We prioritize precision to minimize false positives, given the extreme class imbalance between human-written and AI-generated hate speech in the wild.

Benchmark Construction. We construct an evaluation set of 2K samples, comprising 1K human-written and 1K AI-generated texts, drawn from two publicly available hate speech datasets. The human samples are drawn from the Measuring Hate Speech dataset [15], which contains social media comments with continuous hate speech scores derived from multi-annotator labels. We sample 1,000 hateful samples, providing human-written hate speech stylistically closer to /pol/ than formal benchmarks. The AI-generated samples are drawn from HateBenchSet [43], which contains samples generated by six LLMs, including GPT-3.5, GPT-4, Vicuna, Baichuan2, Dolly2, and OPT, through direct prompting and jailbreaking. We sample 1,000 hateful samples from HateBenchSet, representing the most likely mode of AI-assisted hate speech production on anonymous platforms. Together, the two datasets provide a balanced benchmark covering both human-written and AI-generated hate speech, enabling a fair comparison of detector performance.

Table 2: Benchmark results of AI text detectors.

Method	Acc	Prec	Recall	F1	AUC
Log-Likelihood	0.835	0.817	0.864	0.840	0.919
DetectGPT	0.764	0.776	0.741	0.758	0.835
Entropy	0.672	0.655	0.726	0.689	0.746
Pangram V3	0.680	0.995	0.361	0.530	0.680
OpenAI Detector	0.462	0.474	0.683	0.559	0.453

Benchmark Results and Detector Selection. We evaluate five detection methods on this benchmark. For metric-based methods, i.e., Log-Likelihood [44], Entropy [8], and DetectGPT [24], we use GPT-2 Medium [37] as the base model to compute per-text scores and report AUC as the primary threshold-free metric. Accuracy and F1 are computed at the oracle-optimal threshold using Youden’s J [78] and represent upper-bound performance. For model-based methods, we employ the OpenAI Detector [28], which uses the pre-trained roberta-base-openai-detector model, and the Pangram V3 commercial API [29], both with a default classification threshold of 0.5. The results are summarized in Table 2. While Log-Likelihood yields the highest AUC of 0.919, its precision is limited to 0.817, implying that nearly 20% of flagged posts would be false positives. Applied to tens of thousands of candidate posts, such a margin of error would generate thousands of spurious detections and potentially distort the observed trends in AI-generated content.

We therefore select Pangram V3 [29] for our longitudinal analysis. Although its AUC is moderate (0.680), Pangram achieves near-perfect precision (0.995): when it flags a post as AI-generated, it is almost certainly correct. Furthermore, in the pre-ChatGPT era,³ the detector flagged 8 posts as AI-generated out of a total of 5,239 anti-Indian and anti-Semitic posts. This negligible false-positive rate is consistent with expected model noise and does not undermine the validity of our general conclusions. The trade-off is low recall (0.361), meaning all AI prevalence rates should be interpreted as conservative lower bounds. We further discuss this limitation in Section 7.

Detection Pipeline. We first filter posts exceeding 1,000 characters, as AI text detectors require sufficient text length for reliable classification [24]. This step selects the long-form posts where LLM assistance is most plausible. We then clean each post by removing 4chan-specific formatting such as quote links, URLs, and HTML tags, and normalizing whitespace before submitting it to Pangram. Pangram returns a classification label of AI, AI-Assisted, Human, or Mixed, along with a continuous AI-fraction score between 0 and 1. We apply Pangram to two target groups over the period from January 2022 to January 2026. For anti-Indian posts, we filter posts using slurs pajeet, pajeet, pajeets, poojeet, and pejeet, yielding 17,064 posts for analysis. For anti-Semitic posts, we filter posts with slurs kike, kikes, and heeb, drawing 300 posts per month via stratified sampling for temporal coverage, yielding 14,692 posts at comparable scale.

³We treat posts before November 30, 2022 (ChatGPT’s release date) [66] as the pre-ChatGPT era.

Table 3: AIDE (original weights) per-generator results on AIGIBench. Sorted by accuracy. H.Acc/A.Acc = human-created/AI-generated image accuracy.

Generator	Type	Acc	AP	H.Acc	A.Acc
BLIP	Diff. personalized	99.03	99.98	98.09	99.98
GLIDE	Diff. text-to-image	94.31	97.66	89.00	99.62
SDXL	Diff. text-to-image	93.27	97.69	86.80	99.73
SD3	Diff. text-to-image	92.89	97.19	86.80	98.98
ProGAN	GAN generation	92.74	98.75	88.22	97.25
FaceSwap	GAN face swap	91.09	95.18	87.80	94.45
Imagen3	Diff. text-to-image	90.51	95.65	81.56	99.47
FLUX1-dev	Diff. text-to-image	89.81	96.77	81.78	97.84
StyleGAN-XL	GAN generation	88.29	93.15	82.42	94.16
Midjourney	Diff. text-to-image	74.83	87.72	53.47	96.20
DALL-E 3	Diff. text-to-image	66.27	75.64	81.51	51.02
CommunityAI	In-the-wild	66.17	80.75	94.73	37.60
SocialRF	In-the-wild	62.43	69.25	86.73	38.13

3.5 AI-Generated Image Detection

Existing AI image detectors are primarily evaluated on clean, curated datasets. Their reliability on the compressed, low-quality, and heavily edited images typical of 4chan remains unclear. To identify a detector suitable for our setting, we benchmark five methods and evaluate whether they retain meaningful detection ability on realistic, degraded images.

Benchmarks and Detectors. We evaluate five detectors spanning different technical approaches: AIDE [77], SAFE [19], FerretNet [21], C2P-CLIP [47], and CNNDetection [53]. We test them on two complementary benchmarks. AIGIBench [20] covers 25 generator subsets, including GAN-based generation, diffusion text-to-image models, face swapping, and in-the-wild community content. Chameleon [77] tests a harder scenario with 26K images from professional photographers and AI artists, including various post-processing distortions.

Benchmark Results and Detector Selection. Table 4 reports the results. We distinguish between each detector’s official pretrained weights and, where available, models retrained on AIGIBench’s training split, denoted “S-II.” We report four metrics: overall accuracy (Acc), average precision (AP), human-created image accuracy (H.Acc), and AI-generated image accuracy (A.Acc).

On AIGIBench, AIDE with its original weights performs best overall, with Acc of 85.31% and AP of 91.37%, balancing both H.Acc (83.77%) and A.Acc (86.86%). Per-generator analysis (Table 3) shows that AIDE excels on diffusion models and GANs but degrades on in-the-wild community content (CommunityAI 66.17%, SocialRF 62.43%), the scenario most similar to 4chan’s /pol/.

Other detectors also perform reasonably on AIGIBench, with SAFE, FerretNet, and C2P-CLIP achieving H.Acc between 85.81% and 96.98% and A.Acc between 42.62% and 66.71%. However, on Chameleon, which better approximates the 4chan setting, all detectors except AIDE collapse: their A.Acc drops below 6.00%. AIDE retains the highest AP of 70.09% with A.Acc of 26.00%. Notably, AIDE with its original weights outperforms the Setting-II retrained variant. This is because the original checkpoint was trained on Gen-Image, which covers 8 diverse generators, whereas the Setting-II checkpoint was retrained on only ProGAN and SD-v1.4 across 4

Table 4: Benchmark results of AI image detectors. Wt. = Weights: “S-II” = retrained on AIGIBench training split; “orig” = original pretrained weights. H.Acc = human-created image accuracy (higher means fewer false alarms on human-created images). A.Acc = AI-generated image accuracy (higher means more AI images detected).

Detector	Wt.	AIGIBench				Chameleon			
		Acc	AP	H.Acc	A.Acc	Acc	AP	H.Acc	A.Acc
AIDE	orig	85.31	91.37	83.77	86.86	65.75	70.09	95.62	26.00
AIDE	S-II	77.65	82.72	88.14	67.11	-	-	-	-
SAFE	orig	76.29	78.94	85.81	66.71	59.19	50.64	99.56	5.47
SAFE	S-II	79.67	81.80	96.91	62.34	-	-	-	-
FerretNet	orig	78.15	83.15	91.40	64.83	59.14	43.23	99.25	5.77
C2P-CLIP	orig	69.82	77.39	96.98	42.62	57.59	40.19	99.89	1.32
CNNDet	orig	52.98	63.58	99.76	6.10	56.94	32.99	99.70	0.04

object categories. Since AIGIBench is designed to evaluate cross-generator generalization with 25 generators, most of which are unseen during Setting-II training, the retrained model fails to generalize. We therefore select AIDE with its original weights as our primary detector, as it is the only method that retains usable detection ability on the noisy images found in online communities.

Detection Pipeline. We apply AIDE to images retrieved via CLIP semantic search from our image corpus (Section 3.3). For anti-Indian imagery, we use the query “pajeet,” the primary slur targeting Indians on /pol/. For anti-Semitic imagery, we use “antisemitic caricature,” which captures the most visually distinctive hateful content such as the Happy Merchant meme [16] and related stereotypical depictions. Image detectors produce non-zero detection rates even in 2021, before widely accessible generators such as Stable Diffusion and Midjourney became available. We therefore treat the 2021 detection rate as a conservative baseline, which may reflect a mixture of false positives on compressed and heavily edited 4chan images and a small number of early GAN-generated content. We report absolute detection rates but interpret results in terms of relative changes from the 2021 baseline.

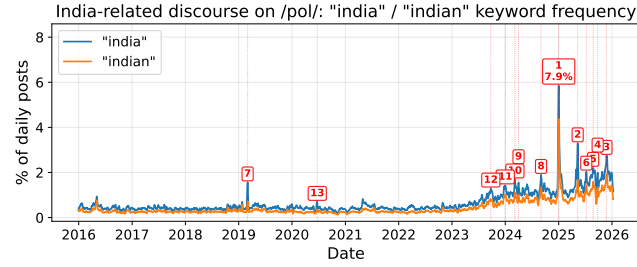
4 RQ1: Temporal Dynamics of Hateful Content

Before examining AI-generated content, we first characterize the landscape of hateful content on /pol/. We analyze temporal trends and peak drivers, discover slurs via Word2Vec, score post toxicity, and track hate imagery through CLIP semantic retrieval.

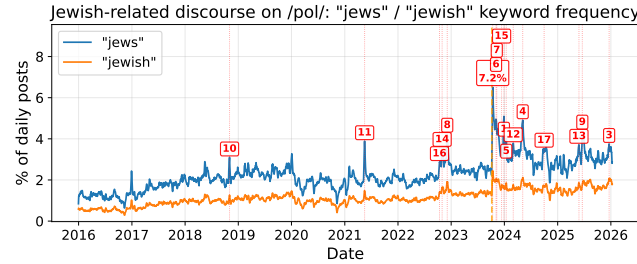
4.1 Temporal Trends and Event Drivers

The two targets exhibit fundamentally different growth dynamics. We detect 13 peaks for India-related and 17 peaks for Jewish-related discussion (Figures 3a-3b; Tables 5-6).

India-Related Discussion: Gradual, Discourse-Driven Escalation. India-related discussion remains low at ~0.3–0.5% of daily posts from 2016 to 2023, with only brief event-driven spikes (2019 Balakot airstrikes, 2020 Galwan Valley clash, 2023 Canada–India crisis) that quickly subside. The sustained escalation above 1% begins in mid-2024, driven not by a single catalytic event but by community-organized campaigns. The pattern shifts in 2024, when the community begins generating its own triggers through



(a) India-related: “india”/“indian.” Circled numbers correspond to events in Table 5.



(b) Jewish-related: “jews”/“jewish.” Oct 7, 2023 (orange dotted line) triggers regime change. Circled numbers correspond to events in Table 6.

Figure 3: Daily percentage of India- and Jewish-related posts on /pol/. [E] = event-driven, [D] = discourse-driven.

discourse-driven campaigns (copypasta raids, organized hate campaigns) that require no external news event. Real-world events can still amplify hate within this active community, as seen when the Atul Subash case intersected with the H-1B visa debate (7.89%). Overall, 6 of 13 peaks (46.15%) are discourse-driven, and all six fall in 2024–2025, whereas every pre-2024 peak is event-driven, indicating that anti-Indian hate on /pol/ has only recently developed self-sustaining community dynamics.

Jewish-Related Discussion: Event-Driven Regime Change. Jewish-related discussion follows a starkly different pattern: it is anchored to real-world events throughout the observation period. Early peaks include the 2018 Pittsburgh synagogue shooting, the 2021 Israel–Palestine crisis, and the 2022 Kanye West antisemitism controversy. The October 7 Hamas attack in 2023 acts as a structural break, producing an immediate spike to 7.24% on Oct 8—3.4 times the pre-attack baseline of 2.12%. The mean frequency of “jews” increased from 2.12% to 3.21%, representing a 1.5-fold increase that persisted as a new elevated baseline. Subsequent escalations in the Gaza war and related events (e.g., Chabad tunnels, UPenn controversy, Israel–Iran conflict) produce corresponding peaks throughout 2024–2025. In total, 16 of 17 peaks (94.12%) are event-driven, with only one discourse-driven peak, a copypasta that recurs each Christmas.

4.2 Slurs and Their Temporal Trends

The trend analysis above measures discussion *volume* using neutral reference terms, but does not reveal whether that discussion is

Table 5: Detected peaks in India-related discussion. [E] refers to event-driven peaks triggered by external news events; [D] means discourse-driven peaks induced by community activity without external triggers.

#	Peak date	%	Width	Trigger
1	2024-12-27	7.89	5.7d	[E] Atul Subash suicide [70] + Musk H-1B debate [68]
2	2025-05-06	4.52	5.3d	[E] India-Pakistan armed conflict [73]
3	2025-11-23	3.78	4.2d	[D] “Fecalwall” hate campaign
4	2025-09-20	3.62	1.6d	[E] Trump H-1B fee + “Operation Clog” campaign [76]
5	2025-08-23	2.67	8.2d	[E] Harjinder Singh case [26]
6	2025-07-06	2.57	4.5d	[D] Organized copypasta raid
7	2019-02-26	2.53	4.1d	[E] India airstrikes on Balakot, Pakistan [56]
8	2024-08-31	2.52	6.6d	[D] Anti-Indian hate wave
9	2024-03-26	2.09	3.3d	[D] GoyimTV-driven anti-Indian hate raid
10	2024-03-03	2.00	6.1d	[D] Anti-Indian sexual copypasta raid
11	2024-01-01	1.79	6.5d	[D] Anti-Indian sexual copypasta raid
12	2023-09-19	1.64	3.0d	[E] Canada-India diplomatic crisis [61]
13	2020-06-16	1.50	1.7d	[E] India-China Galwan Valley clash [57]

Table 6: Detected peaks in Jewish-related discussion. [E] refers to event-driven peaks triggered by external news events; [D] means discourse-driven peaks induced by community activity without external triggers.

#	Peak date	%	Width	Trigger
1	2023-12-24	8.93	2.5d	[D] Antisemitic child abuse conspiracy
2	2023-10-08	7.24	8.5d	[E] Oct 7 Hamas attack [64]
3	2025-12-14	5.75	1.6d	[E] 2025 Bondi Beach shooting [71]
4	2024-05-01	5.71	15.5d	[E] Columbia University protests [67]
5	2024-01-09	5.60	1.8d	[E] Chabad tunnels controversy [4, 63]
6	2023-11-01	5.22	9.6d	[E] Jabalia camp strikes [62]
7	2023-11-17	4.99	3.1d	[E] Al-Shifa raid [60]; bin Laden letter [32]
8	2022-12-02	4.95	9.8d	[E] Kanye posts swastika; Twitter ban [59]
9	2025-06-25	4.75	17.6d	[E] Twelve-Day War ceasefire [74]
10	2018-10-27	4.71	3.8d	[E] Pittsburgh synagogue shooting [55]
11	2021-05-12	4.42	7.6d	[E] 2021 Israel-Palestine crisis [58]
12	2024-02-27	4.32	4.7d	[E] Self-immolation of Aaron Bushnell [69]
13	2025-05-22	4.29	4.7d	[E] 2025 Capital Jewish Museum shooting [72]
14	2022-11-04	4.16	4.7d	[E] Kyrie Irving suspension [25]
15	2023-12-09	4.12	2.3d	[E] UPenn president resigns over antisemitism [31]
16	2022-10-09	4.03	1.2d	[E] Kanye “death con 3” tweet [59]
17	2024-09-28	3.98	34.7d	[E] Assassination of Nasrallah [65]

hateful. We now turn to the hate vocabulary itself, as presented in Table 1. Columns 1–2 show the top-15 nearest neighbors of each neutral reference term, revealing how the /pol/ community semantically associates “india” and “jews” with other words. Column 3 shows the most similar words to the embedding obtained by adding a hate seed’s embedding to a reference term’s embedding, surfacing terms that are simultaneously close to the target group and to hateful semantics.

Three findings stand out. First, the nearest neighbors of “india” already include hateful coinages: variants of the slur “pajeet” [54] (“pajeetland,” “poojeetistan,” “pooindia”), showing that even the neutral country name has become semantically entangled with slurs on /pol/. Second, the vector arithmetic column reveals the

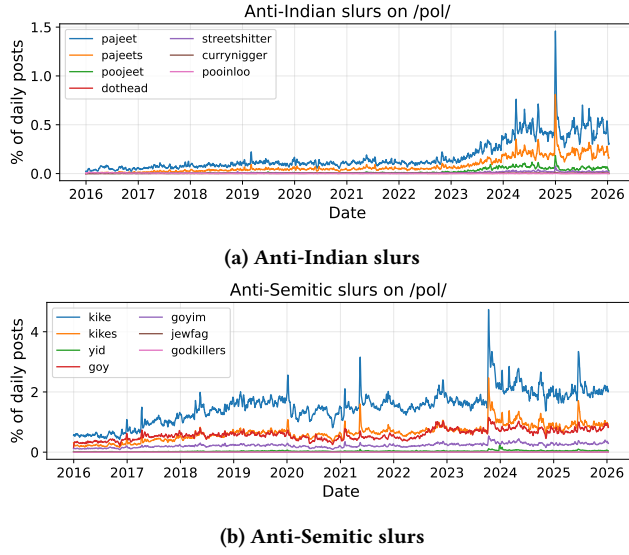


Figure 4: Daily frequency of hate slurs on /pol/. Anti-Indian slurs show explosive growth in 2024–2025; anti-Semitic slurs show event-driven spikes on a higher baseline.

dominant anti-Indian hate narratives: hygiene stereotypes (“pooinloo” [17], “streetshitter” [75]), compound racial slurs (“currynigger,” “dothead”), and hateful country names (“poojetistan,” “shit-skindia”). Third, the anti-Semitic columns confirm that established slurs (“kike,” “goy”) dominate.

We track the daily frequency of slurs for both target groups. For anti-Indian content, we monitor: “pajeet,” “pajeets,” “poojeet,” “dothead,” “streetshitter,” “currynigger,” and “pooinloo,” combining slurs from Word2Vec discovery (Table 1) and external documentation [6]. For anti-Semitic content, we monitor “kike,” “kikes,” “yid,” “goy,” “goyim,” “jewfag,” and “godkillers.” Figure 4 displays the trends for the most prominent slurs from each group.

Anti-Indian Slur Vocabulary Is Actively Expanding. Anti-Indian slurs grow in both usage and variety (Figure 4a). “pajeet” rises from ~0.2% of daily posts in 2020 to a peak of 1.5% in 2025, a 7-fold increase. The derived slur “poojeet” emerges from near-zero in 2024, indicating active slur vocabulary expansion. This growth trajectory closely mirrors the neutral keyword trends in Figure 3a, suggesting that rising discussion volume and slur usage reinforce each other.

Anti-Semitic Slurs Draw on an Established Vocabulary. Anti-Semitic slurs show a different pattern (Figure 4b). “Kike” and “kikes,” both surfaced in Table 1, maintain a persistent high baseline of ~1% with event-driven spikes, most notably around Oct 7, 2023, aligning with the same event-driven peaks observed for neutral reference terms “jews” and “jewish” in Figure 3b. Established terms such as “yid” and “goy” show steady usage, while newer compound slurs discovered by our method, including “jewfag” and “godkillers,” remain at trace levels. This contrast suggests that anti-Semitic vocabulary on /pol/ is comparatively stable, relying on well-established slurs rather than actively coining new ones.

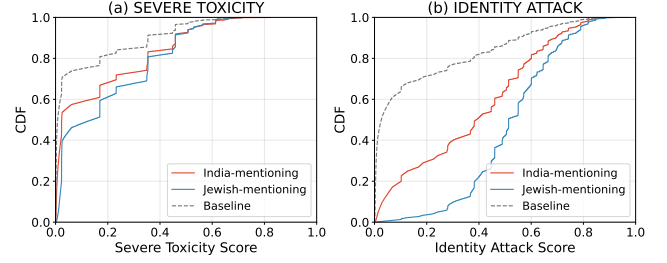


Figure 5: CDFs of (a) SEVERE TOXICITY and (b) IDENTITY ATTACK scores (Perspective API).

Table 7: CLIP semantic image search results. Rates are percentages of images posted that week.

Target	Query	2021	2025	Trend
Anti-Indian	india	0.046%	0.207%	×4.5↑
	pajeet	0.070%	0.233%	×3.3↑
	anti indian meme	0.067%	0.222%	×3.3↑
	street shitter	0.047%	0.047%	flat
Anti-Semitic	jewish	0.183%	0.246%	×1.3↑
	antisemitic caricature	0.153%	0.215%	×1.4↑
	jew conspiracy	0.170%	0.241%	×1.4↑
	anti jewish meme	0.195%	0.279%	×1.4↑

4.3 Toxicity Analysis

The slur analysis above reveals a growing hate vocabulary, but hate speech extends beyond explicit slurs. Using the toxicity scoring methodology described in Section 3.2, Figure 5 compares the three groups. Both groups are substantially more toxic than the baseline across both dimensions. The effect is especially pronounced for IDENTITY ATTACK: 58.04% of posts mentioning Jews and 35.87% of posts mentioning India score above 0.5, compared to only 13.12% of baseline posts. Posts mentioning Jews are the most toxic overall, likely because “jews” appears overwhelmingly in hostile contexts on /pol/, whereas “india” also appears in neutral geopolitical discussions (e.g., trade, elections). For SEVERE TOXICITY, the gap is narrower but still clear: 7.88% (Jews) and 7.38% (India) exceed 0.5, versus 3.14% for the baseline.

4.4 Hate Imagery Trends

We now turn to the visual modality. Table 7 and Figure 6 present the weekly rate of images retrieved by each query, normalized by total posting volume on /pol/ (see Section 3.3). India-related imagery (Figure 6a) shows sustained growth: “pajeet” rates more than tripled from 0.070% in 2021 to 0.233% in 2025, consistent with the textual slur explosion (Figure 4). “Street shitter” imagery remained flat (0.047% in both periods), suggesting a vocabulary shift toward the “pajeet” meme template. Jewish-related imagery (Figure 6b) shows a higher but stable baseline with event-driven spikes. The most prominent spike occurs on October 7, 2023: “antisemitic caricature” retrievals surge to 0.562%, roughly 3.7 times the 2021 average. All four Jewish-related queries exhibit elevated rates during this week, confirming a reactive surge following the Hamas attack. This cross-modal alignment reinforces findings on the text level: both

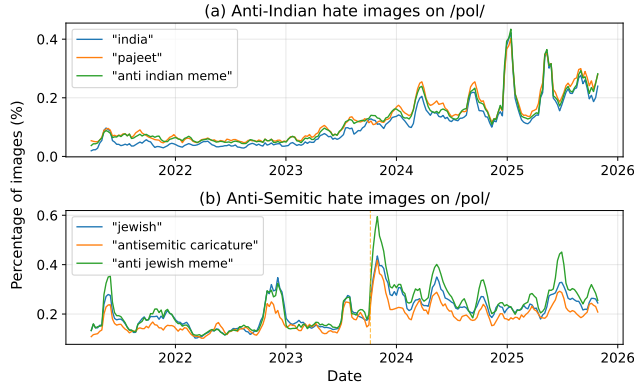


Figure 6: Weekly rate of hate imagery on /pol/ (% of images posted), retrieved via CLIP semantic search across 16.9M images. (a) Anti-Indian imagery shows sustained growth in 2024–2025. (b) Anti-Semitic imagery spikes on Oct 7, 2023 (orange dotted line).

modalities exhibit similar contrasting dynamics, i.e., gradual growth for India, event-driven spikes for Jews, strengthening the evidence that the observed patterns are genuine and not due to limitations of any single measurement method.

Take-Away (RQ1): Hate against both target groups is growing, but through fundamentally different mechanisms. Anti-Indian hate is emerging and self-sustaining: 46.15% of peaks are driven by organized community campaigns, slur vocabulary is expanding explosively, and hateful imagery rates tripled from 0.070% to 0.233% between 2021 and 2025. Anti-Semitic hate is established and reactive: 94.12% of peaks are tied to external events, with October 7 acting as a structural break, a stable slur vocabulary, and image rates spiking 3.7-fold at October 7. These contrasting dynamics, confirmed across text, images, and toxicity, demand different moderation strategies.

5 RQ2: AI-Generated Content in Hate Discourse

We now examine whether AI-generated content has emerged in these communities.

5.1 AI-Generated Text

We apply the Pangram detection pipeline from Section 3.4 to 17,064 anti-Indian and 14,692 anti-Semitic long-form posts spanning January 2022 to January 2026. Table 8 reports quarterly detection rates and Figure 7 shows the monthly trend.

AI-Generated Text Was Essentially Absent Before ChatGPT. Among 1,947 anti-Indian posts before ChatGPT's launch in November 2022, Pangram flagged zero as AI-generated. For anti-Semitic posts, 8 of 3,292 pre-ChatGPT posts were flagged (0.24%), possibly reflecting early usage of small language models.

AI-Generated Text Has Grown Steadily Since ChatGPT's Release. Detection rates rose steadily from 2023 onward. By 2025, the AI rate reached 1.42% for anti-Indian and 1.28% for anti-Semitic posts, both significantly above 2023 levels (two-proportion z-test:

Table 8: Quarterly AI-generated text detection rates for anti-Indian and anti-Semitic hateful posts.

Quarter	Anti-Indian		Anti-Semitic	
	# Posts	AI Rate	# Posts	AI Rate
2022-Q1	704	0.00%	900	0.00%
2022-Q2	395	0.00%	900	0.44%
2022-Q3	498	0.00%	898	0.22%
2022-Q4	524	0.19%	899	0.22%
2023-Q1	505	1.58%	900	0.33%
2023-Q2	595	0.67%	900	0.56%
2023-Q3	810	0.74%	899	0.33%
2023-Q4	830	0.12%	899	0.89%
2024-Q1	1,496	0.47%	899	0.22%
2024-Q2	1,399	1.07%	900	0.33%
2024-Q3	1,538	0.65%	900	0.22%
2024-Q4	1,313	0.84%	900	0.56%
2025-Q1	1,173	0.51%	899	1.11%
2025-Q2	1,150	0.87%	899	2.11%
2025-Q3	1,888	1.69%	900	1.11%
2025-Q4	1,843	2.06%	900	0.78%
2026-Q1	403	0.25%	300	1.00%
Total	17,064	0.88%	14,692	0.60%

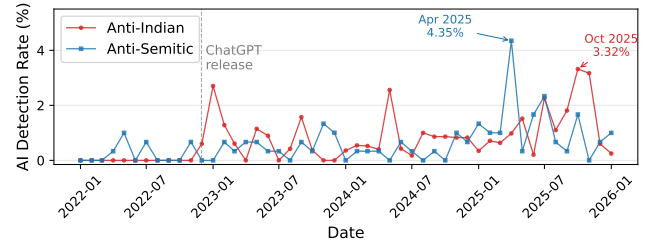


Figure 7: Monthly AI-generated text detection rates for anti-Indian and anti-Semitic hateful posts, January 2022 to January 2026. The dashed line marks ChatGPT's release.

$z=2.91$, $p=0.002$ and $z=3.36$, $p<0.001$). Every flagged post received an AI-fraction score of 1.0, indicating fully LLM-generated text rather than human-AI co-writing. This acceleration coincides with the spread of open-source models such as Llama 3, Mistral, and Qwen that can run locally without content safety filters.

AI-Generated Posts Are Longer and More Toxic. AI-generated posts are significantly longer than human-written ones: 1,501 vs. 1,321 characters for anti-Indian posts (Mann-Whitney $p<0.0001$, rank-biserial $r=0.25$) and 1,621 vs. 1,392 for anti-Semitic posts ($p<0.0001$, $r=0.39$), consistent with the verbose output typical of LLMs. Cross-referencing with Perspective API shows that AI-generated anti-Indian posts are also more toxic, with higher SEVERE TOXICITY (mean 0.41 vs. 0.26, Mann-Whitney $p<0.0001$), though IDENTITY ATTACK scores do not differ significantly (0.47 vs. 0.47, $p=0.07$). These results suggest that LLMs are used to produce content that is not only hateful but more extreme than what human users typically write.

Table 9: Yearly AI-generated image detection results for anti-Indian and anti-Semitic hate imagery. # images refers to the number of images retrieved via CLIP semantic search for each target group. Non-zero pre-2022 rates may reflect potential usage of early GAN-based models.

Year	# images		AIDE >0.5		AIDE >0.7	
	Indian	Jewish	Indian	Jewish	Indian	Jewish
2021	2,025	4,461	10.72%	9.19%	2.81%	2.80%
2022	2,303	6,082	10.68%	11.10%	2.95%	3.52%
2023	3,576	7,179	10.18%	12.68%	2.74%	3.83%
2024	6,085	6,881	11.42%	11.77%	2.78%	3.58%
2025	5,428	5,138	11.15%	12.65%	3.02%	3.35%

AI-Generated Posts Receive Less Engagement. Despite their higher toxicity, AI-generated posts attract fewer replies. For anti-Indian posts, the median reply count after an AI-generated post is 14, compared to 46 for human posts (Mann-Whitney $p < 0.0001$). For anti-Semitic posts, no significant difference is observed (median 53 for both, $p = 0.98$). The lower engagement may be because AI-generated text tends to be more formulaic and repetitive [41], making it less engaging for readers.

AI-Generated Text Spikes During Geopolitical Flashpoints. AI-generated content does not appear at a constant rate but spikes during geopolitical flashpoints. Anti-Indian AI rates peaked in late 2025, reaching 3.32% in October and 3.17% in November, more than three times the overall average. For anti-Semitic posts, the largest spike occurred in April 2025 at 4.35%, over seven times the average. July 2025 is the only month where both groups simultaneously exceeded 2%, suggesting that major geopolitical events can trigger AI-assisted hate campaigns across multiple target communities.

5.2 AI-Generated Images

We apply AIDE (Section 3.5) to images retrieved via CLIP semantic search for two hate targets: “pajeet” for anti-Indian imagery, covering 19,540 images, and “antisemitic caricature” for anti-Semitic imagery, covering 30,139 images, both spanning 2021–2025. Table 9 reports the yearly detection rates.

Image AI Rates Must Be Read as Relative Change from 2021. Unlike AI-generated text detection, where the pre-ChatGPT AI rate was near zero, AI-generated image detectors flag a non-trivial fraction of images even in 2021, before widely accessible generators such as Stable Diffusion and Midjourney. These baseline detections may reflect potential usage of early GAN-based models. We therefore interpret image results in terms of relative change from the 2021 baseline rather than absolute prevalence.

AI Detection Rates Rise for Anti-Semitic Imagery but Not Anti-Indian. Both targets show higher detection rates in 2025 than in 2021, but only the anti-Semitic increase is statistically significant. We test each series with the Cochran–Armitage trend test across the five annual bins. At the default threshold of 0.5, anti-Indian imagery rises from 10.72% to 11.15% ($z = 1.16$, $p = 0.25$), and anti-Semitic imagery from 9.19% to 12.65% ($z = 5.03$, $p < 0.001$), with a spike in 2023Q4 coinciding with the post-October 7 surge in antisemitic content volume documented in RQ1. At the conservative threshold of 0.7, the increases are smaller: 2.81% to 3.02% for anti-Indian and

2.80% to 3.35% for anti-Semitic imagery, a much smaller absolute change that does not reach significance ($p = 0.69$ and $p = 0.22$). However, the high false-positive baseline complicates interpretation. AIDE already flags approximately 10% of images in 2021, before widely accessible text-to-image generators existed, meaning the true AI signal is embedded in substantial detector noise. A real increase from 0% to 3% in actual AI-generated content, for example, would appear as a rise from 10% to 13% in detection rates, which is difficult to distinguish from normal fluctuation. The observed trends are therefore consistent with a gradual adoption of AI-generated imagery, but we cannot quantify the true rate with confidence given current detector limitations.

AI-Flagged Hate Imagery Has Tripled in Absolute Volume.

Regardless of whether the AI rate is rising, the absolute number of AI-flagged hate imagery has grown substantially because the underlying hateful content volume has expanded. Pajeet-related images increased from 2,025 in 2021 to 6,085 in 2024, a 3.0-fold increase that tracks the hate discourse growth documented in RQ1. At a detection rate of 10–11%, AIDE flags 217 images in 2021 and 695 in 2024.

Detecting AI-Generated Images Is Harder Than Detecting AI-Generated Text. AI-generated image detection on 4chan’s /pol/ presents a fundamentally harder problem than text detection. Text detection has a clean zero-baseline in pre-ChatGPT 2022 and shows clear acceleration from 0.00% to 1.42%, doubling year-over-year by 2025. Image detection lacks this clean baseline, as generative models like GANs had already established a presence long before this period. Generating convincing hate imagery also requires greater technical effort than prompting an LLM for text, which may further explain the slower and less visible adoption.

Take-Away (RQ2): AI-generated text has infiltrated hate communities: absent before ChatGPT, it now reaches 1.42% of long-form posts and is growing year-over-year. AI-generated images show a slight upward trend for anti-Semitic imagery, from 9.19% to 12.65%, but anti-Indian imagery remains flat at 10–11%.

6 RQ3: Co-occurrence and Amplification Effects of AI-Generated Content

RQ1 showed that hateful content is growing online, and RQ2 showed that AI-generated content is emerging within it. This section asks whether the two trends are linked or merely coincidental. We test this through four progressively stronger analyses and compare anti-Indian with anti-Semitic hate to assess generalizability.

Correlation Between AI-Generated Content and Hate Intensity. We ask whether higher rates of AI-generated content also show more intense hate activity. We compute the Spearman rank correlation [45] between the monthly AI detection rate and the monthly frequency of the slur “pajeet” across 49 months of Pangram detection data (2022–2026). As shown in Table 10, the correlation is strong and significant for anti-Indian hate ($\rho = 0.55$ – 0.57 , $p < 0.0001$): months with heavier slur usage show higher AI detection rates. Figure 8 illustrates this visually: both slur frequency and AI rate emerge after ChatGPT’s release and rise together through 2025. For

Table 10: Spearman correlations between monthly slur frequency and AI-generated content metrics ($N=49$ months, 2022–2026).

Target	Metric Pair	ρ	p -value
Anti-Indian	"pajeet" freq. vs. Pangram AI rate	0.57	<0.0001
	"pajeet" freq. vs. fraction_ai>0.9 rate	0.57	<0.0001
	"pajeet" freq. vs. mean fraction_ai	0.55	<0.0001
Anti-Semitic	"kike" freq. vs. Pangram AI rate	0.21	0.15
	"kike" freq. vs. fraction_ai>0.9 rate	0.21	0.15
	"kike" freq. vs. mean fraction_ai	0.19	0.20

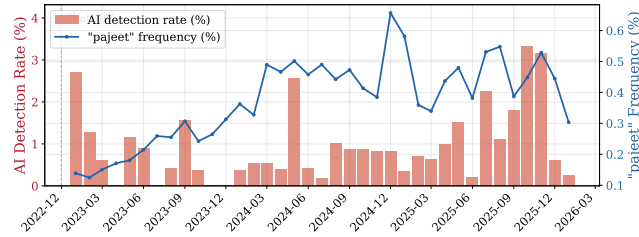


Figure 8: Monthly anti-Indian slur frequency (blue line, right axis) and AI detection rate (red bars, left axis). Both trends emerge after ChatGPT's release (dashed line) and rise in tandem through 2025.

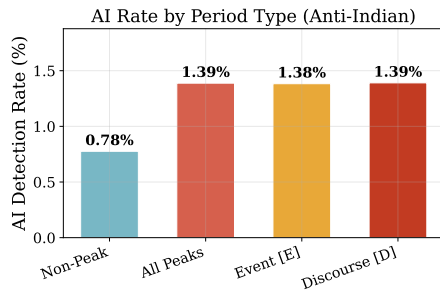


Figure 9: Anti-Indian AI detection rate by period type. Peak months show nearly double the AI rate of non-peak months.

anti-Semitic hate, the same analysis yields no significant correlation ($\rho=0.19$ – 0.21 , $p>0.15$). AI-generated content and hate activity move together for anti-Indian hate but not for anti-Semitic hate.

AI Concentration at Hate Peaks. To test whether AI-generated content clusters around hate peaks, we label each month as a peak month if it contains one of the 13 detected peaks (Table 5) and compare AI rates between peak and non-peak periods with a two-proportion z-test over post-level counts. For anti-Indian hate, the AI rate during peak months is nearly double that of non-peak months (1.39% vs. 0.78%; $z=3.62$, $p<0.001$; Figure 9). For anti-Semitic hate, AI rates remain similar between peak and non-peak months (0.60% vs. 0.77%, $p=0.36$).

AI-Generated Posts Are More Toxic Than Human-Written Ones. To test whether AI-generated posts are more harmful, we

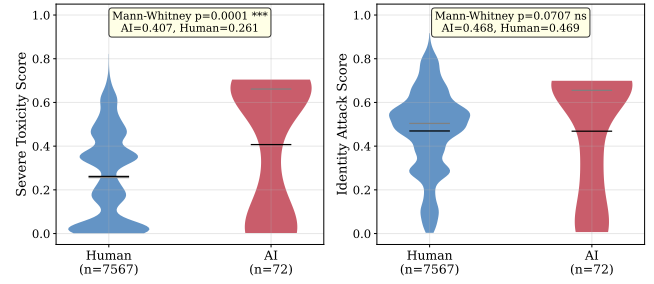


Figure 10: Toxicity distributions of AI-generated vs. human-written anti-Indian hateful posts. AI-generated posts score significantly higher on SEVERE TOXICITY ($p<0.0001$) and concentrate in the high-toxicity range.

Table 11: Summary of RQ3 analyses across hate targets.

Analysis	Anti-Indian	Anti-Semitic
Temporal correlation	$\rho=0.57$, $p<0.0001$	$\rho=0.21$, $p=0.15$
Peak AI rate	1.39% vs. 0.78%	0.60% vs. 0.77%
SEVERE TOXICITY (AI vs. Human)	0.41 vs. 0.26, $p<0.0001$	too few AI samples
Lead-lag peak	lag +3, $\rho=0.35$, $p=0.04$	no significant lag

link Pangram AI labels with the Perspective toxicity scores. For anti-Indian hate, this yields 72 AI-generated and 7,567 human-written anti-Indian hateful posts. As shown in Figure 10, AI-generated posts score 55.72% higher on SEVERE TOXICITY than human posts. The violin plot shows that AI-generated posts cluster in the high-toxicity range, whereas human posts span a much wider distribution. Each AI-generated post therefore contributes more to the overall harmfulness of hate discourse, even though AI-generated posts account for only 1–2% of long-form posts. For anti-Semitic posts, only 16 AI-generated posts overlap with Perspective scores, too few to draw reliable conclusions.

Temporal Direction: Does Hate Lead AI, or Vice Versa? The analyses above show that AI and hate co-occur, but do not reveal which one leads. To explore this, we added a post-hoc lead-lag analysis. We shift the two monthly series relative to each other by -6 to $+6$ months and compute Spearman correlations at each offset. For example, lag +3 correlates each month's hate frequency with the AI rate three months later. As shown in Figure 11, the strongest correlation for anti-Indian hate occurs at lag +3 months: a surge in slur usage predicts a rise in AI-generated content roughly three months later. This is consistent with a reactive adoption pattern in which users first engage in hate discourse organically and then turn to AI tools to scale up content production. A negative correlation at lag -5 is marginally significant but lacks a plausible causal mechanism and relies on only 32 data pairs. For anti-Semitic hate, no significant positive lag is observed.

Summary and Interpretation. Table 11 summarizes the four analyses for both targets. For anti-Indian hate, all four analyses point in the same direction: AI-generated content co-occurs with hate intensity, concentrates during hate peaks, is more toxic than human-written content, and temporally follows hate surges rather than preceding them. For anti-Semitic hate, none of these patterns

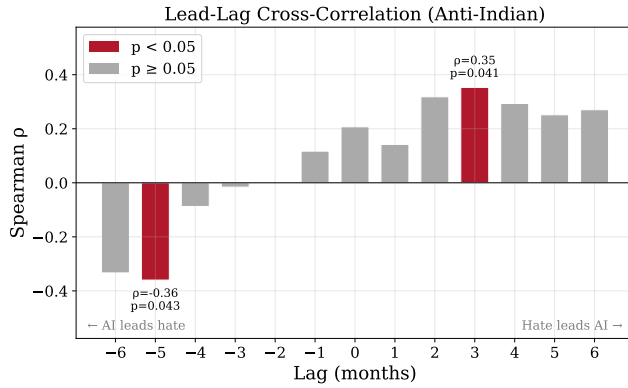


Figure 11: Lead-lag cross-correlation between monthly anti-Indian slur frequency and AI detection rate. The peak at lag +3 ($\rho=0.35$, $p=0.04$) indicates that hate activity precedes AI adoption by about three months. Red bars indicate unadjusted p -values below 0.05.

hold. This contrast is consistent with the different dynamics of the two targets. Anti-Indian hate is a rapidly growing phenomenon fueled by organized community campaigns (Section 4), and these campaigns are precisely the context in which users adopt AI tools for content production. Anti-Semitic hate, by contrast, has a long-established presence on /pol/ with a stable vocabulary and event-driven dynamics, leaving less incentive to adopt new production tools. We note two important caveats. First, we cannot establish causation from observational data; the observed correlations may partly reflect shared external drivers such as news events. Second, the detected 1-2% AI rate is a strict lower bound, since Pangram’s recall is moderate and we mainly analyze lengthy posts. Despite these limitations, the converging evidence supports the conclusion that generative AI is being selectively deployed to amplify emerging, campaign-driven hate speech on /pol/.

Take-Away (RQ3): AI is selectively used to amplify hate towards target groups. For anti-Indian hate, AI-generated content correlates with hate intensity ($\rho=0.57$), concentrates during hate peaks (1.39% vs. 0.78%), is 55.72% more toxic, and follows hate surges by approximately three months. Anti-Semitic hate shows none of these patterns, indicating that AI amplification is topic-specific and concentrated in campaign-driven hate rather than established hate discourse.

7 Discussion

Cross-Validation of the Amplification Effect. The conclusion that AI amplifies anti-Indian hate does not rest on any single analysis. We validate this from four angles: AI-generated content correlates with hate intensity over time, concentrates during hate peaks, is more toxic than human-written content, and temporally follows rather than precedes hate surges. Although each analysis has distinct limitations, their directionally consistent results support the interpretation that AI use is concentrated in emerging, campaign-driven anti-Indian hate. The cross-topic comparison

with anti-Semitic hate provides a useful contrast: applying the same analyses to another hate target yields no significant associations, suggesting that the anti-Indian results are not solely explained by a general platform-wide trend in AI adoption.

Emerging vs. Established Hate. Studying anti-Indian and anti-Semitic hate in parallel reveals two fundamentally different dynamics on the same platform. Anti-Semitic hate on /pol/ has a persistent high baseline, spikes around identifiable external events (94.12% event-driven peaks), and relies on a stable slur vocabulary (“kike,” “goy”) that changes little over time. Platforms can anticipate these surges around predictable triggers such as Israel-Palestine escalations. Anti-Indian hate, by contrast, grows gradually through community-organized campaigns (46.15% discourse-driven peaks) and shows rapid slur growth in both usage and variety (“pajeet” frequency increased 7-fold from 2020 to 2025, with new terms such as “poojeet” and “dothead” appearing in 2024). This pattern is harder to moderate because the activity is internally generated and the vocabulary evolves faster than static keyword filters can track. These contrasting dynamics also explain the RQ3 finding that AI amplification is topic-specific. Anti-Indian hate campaigns are organized efforts that benefit from AI-assisted content production. Anti-Semitic hate draws on decades of established rhetoric, leaving less incentive to adopt new tools.

The Role of AI. Our findings suggest that generative AI amplifies existing hate campaigns rather than initiating new ones. For text, AI-generated content concentrates during anti-Indian hate peaks (1.39% vs. 0.78% in non-peak months), AI-generated posts are 55.72% more toxic than human posts, and the lead-lag analysis shows that hate surges precede AI adoption by about three months. Together, these results suggest that users first engage in hate campaigns organically and then turn to AI tools to scale up content production. The 2025 acceleration in AI rates coincides with the spread of open-source models that can run locally without content safety filters. For images, the pattern differs. AI-generated images appear in hateful content at a roughly constant rate (10–12%), and hate-related images are actually less likely to be AI-generated than typical /pol/ images. Three factors may explain this: generating targeted hate imagery requires bypassing safety filters in most image generators; 4chan’s meme culture favors crude, low-effort aesthetics over polished AI output; and users tend to remix existing images rather than generate new ones.

Implications for AI Safety and Content Governance. Our measurements reveal that safeguard evasion has a downstream consequence. AI-generated hateful content reaches real platforms, concentrates during hate peaks, and is significantly more toxic than human-written posts, making provenance detection operationally valuable. This constitutes the empirical evidence that safeguard evasion, previously demonstrated only in controlled settings, now operates at scale with measurable amplification of harm. The burden thus falls on platform-side moderation rather than model-side safeguards alone. However, current AI detectors recover only a minority of AI-generated content, 36.1% for text (Section 3.4) and 26.00% for images under realistic conditions (Section 3.5), limiting platforms’ ability to attribute hate campaign sources. Beyond provenance, emerging hate targets face a distinct challenge. Static keyword filters degrade as communities coin new slurs, requiring platforms to continuously update detection rules. The Word2Vec

procedure in Section 3.2 can be rerun periodically to surface such coinages from the platform's own corpus.

Limitations. Our paper has limitations. First, for AI-generated content, we mainly analyze posts longer than 1,000 characters due to the limitations of detectors. The AI generation rate among shorter posts remains unknown. Second, confounding factors such as H-1B policy debates and India-Canada diplomatic tensions drive hate sentiment independently of AI availability. Our correlational and lead-lag evidence is suggestive of directionality but does not establish causation. Third, the event-vs-discourse peak classification and the lead-lag analysis were added after data inspection and are exploratory rather than confirmatory. The reported lead-lag peak at lag +3 months is the highest positive correlation across 13 evaluated lags (−6 to +6 months). Its p -value is not adjusted for multiple comparisons. Accordingly, the resulting peak-type comparisons are descriptive, and the lead-lag result should be interpreted as an association rather than evidence of causality. Fourth, Pangram achieves low recall (0.361), missing approximately two-thirds of AI-generated content, and even the best-performing detector we benchmarked does not reach perfect precision. Consequently, all AI prevalence rates reported should be interpreted as strict lower bounds; actual AI usage may be substantially higher. Finally, our study focuses on 4chan's /pol/, an unmoderated fringe platform. Results may not generalize to mainstream platforms with content moderation, where both hate speech volume and AI usage patterns may differ.

8 Conclusion

This paper draws on 465 million text posts and 16.9 million images from 4chan's /pol/ board to study how generative AI interacts with hate speech. We compare two hate targets: anti-Indian hate, which is rapidly growing and driven by organized campaigns, and anti-Semitic hate, which has a stable baseline and responds mainly to external events. This study reveals that AI-generated hateful text has grown to 1.42% of long-form posts by 2025. Second, for anti-Indian hate, AI-generated content correlates with hate intensity, concentrates during hate peaks, is 55.72% more toxic than human-written posts, and follows hate surges by about three months. Third, anti-Semitic hate shows none of these patterns, indicating that AI amplification is selective and concentrated in emerging, campaign-driven hate. These results suggest three priorities for future work: developing AI detectors for the short texts typical of social media, building automated systems to track new slurs as they spread across platforms, and monitoring AI tool adoption on extremist platforms as open-source models make uncensored content generation increasingly easy.

References

- [1] American Jewish Committee. 2024. Translate Hate: AJC's Glossary of Antisemitic Terms, Phrases, Conspiracies, Cartoons, Themes, and Memes. https://www.ajc.org/sites/default/files/pdf/2024-02/AJC_Translate-Hate-Glossary-2.2024.pdf.
- [2] Anti-Defamation League. 2025. Innovative AI Video Generators Produce Antisemitic, Hateful and Violent Outputs. <https://www.adl.org/resources/article/innovative-ai-video-generators-produce-antisemitic-hateful-and-violent-outputs>.
- [3] Zhongjie Ba, Jieming Zhong, Jiachen Lei, Peng Cheng, Qinglong Wang, Zhan Qin, Zhibo Wang, and Kui Ren. 2024. SurrogatePrompt: Bypassing the Safety Filter of Text-to-Image Models via Substitution. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 1166–1180.
- [4] CBS News New York. 2024. Arrests made after brawl with police at Chabad Lubavitch synagogue in Brooklyn. <https://www.cbsnews.com/newyork/news/arrests-made-after-brawl-with-police-at-chabad-lubavitch-synagogue-in-brooklyn/>.
- [5] Center for the Study of Organized Hate. 2025. AI-Generated Hate in India. <https://www.csohate.org/2025/09/29/ai-generated-hate-in-india/>.
- [6] Center for the Study of Organized Hate. 2025. Anti-Indian Racism on X (July–September 2025). <https://www.csohate.org/2025/09/16/anti-indian-racism-on-x-july-september-2025/>.
- [7] Bradley Emi and Max Spero. 2024. Technical Report on the Pangram AI-Generated Text Classifier. *CoRR abs/2402.14873* (2024).
- [8] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 111–116.
- [9] Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y. Zhao. 2024. Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 4822–4836.
- [10] Abhimanyu Hans, Avi Schwarzschild, Valerii Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. In *International Conference on Machine Learning (ICML)*. PMLR, 17519–17537.
- [11] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 3309–3326.
- [12] Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.
- [13] Gabriel Emile Hine, Jeremiah Onalapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *International Conference on Web and Social Media (ICWSM)*. AAAI, 92–101.
- [14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale Similarity Search with GPUs. *IEEE Transactions on Big Data* (2021).
- [15] Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *CoRR abs/2009.10277* (2020).
- [16] Know Your Meme. 2001. Happy Merchant. <https://knowyourmeme.com/memes/happy-merchant>.
- [17] Know Your Meme. 2015. Designated Shitting Streets / Poo in the Loo. <https://knowyourmeme.com/memes/designated-shitting-streets-poo-in-the-loo>.
- [18] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 3197–3207.
- [19] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. 2025. Improving Synthetic Image Detection Towards Generalization: An Image Transformation Perspective. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2405–2414.
- [20] Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. 2025. Is Artificial Intelligence Generated Image Detection a Solved Problem? *CoRR abs/2505.12335* (2025).
- [21] Shuqiao Liang, Jian Liu, Renzhang Chen, and Quanlong Guan. 2025. Ferret-Net: Efficient Synthetic Image Detection via Local Pixel Dependencies. *CoRR abs/2509.20890* (2025).
- [22] Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024. Making Them Ask and Answer: Jailbreaking Large Language Models in Few Queries via Disguise and Reconstruction. In *USENIX Security Symposium (USENIX Security)*. USENIX, 4711–4728.
- [23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations (ICLR)*.
- [24] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning (ICML)*. PMLR, 24950–24962.
- [25] BBC News. 2022. Kyrie Irving suspended over anti-Semitic posts. <https://www.bbc.com/news/world-us-canada-63509715>.
- [26] CBS News. 2025. California trucker accused of causing deadly crash extradited to Florida. <https://www.cbsnews.com/sacramento/news/stockton-trucker-to-be-extradited-florida-deadly-crash/>.
- [27] OpenAI. 2023. GPT-4 Technical Report. *CoRR abs/2303.08774* (2023).
- [28] OpenAI Community. 2019. roberta-base-openai-detector. <https://huggingface.co/openai-community/roberta-base-openai-detector>.
- [29] Pangram Labs. 2025. Pangram AI Detection API v3. <https://docs.pangram.com/api-reference/ai-detection>.

- [30] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *International Conference on Web and Social Media (ICWSM)*. AAAI, 885–894.
- [31] PBS NewsHour. 2023. Liz Magill, UPenn president, resigns after antisemitism testimony draws backlash. <https://www.pbs.org/newshour/education/liz-magill-upenn-president-resigns-after-antisemitism-testimony-draws-backlash>.
- [32] The Washington Post. 2023. How Osama bin Laden's 'Letter to America' reached millions online. <https://www.washingtonpost.com/style/2023/11/16/guardian-osama-bin-laden-letter-to-america/>.
- [33] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. In *International Conference on Learning Representations (ICLR)*.
- [34] Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE.
- [35] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.
- [37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* (2019).
- [38] Radim Rehrák and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *International Conference on Language Resources and Evaluation (LREC)*. ELRA, 46–50.
- [39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 10684–10695.
- [40] Secure Community Network. 2025. New Report Warns Extremist Groups Increasingly Using AI to Intensify Antisemitic Propaganda. <https://www.securecommunitynetwork.org/articles/cbs-news-new-report-warns-extremist-groups-increasingly-using-ai-to-intensify-antisemitic-propaganda/>.
- [41] Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024. Detection and Measurement of Syntactic Templates in Generated Text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 6416–6431.
- [42] Xinyue Shen, Xinlei He, Michael Backes, Jeremy Blackburn, Savvas Zannettou, and Yang Zhang. 2022. On Xing Tian and the Perseverance of Anti-China Sentiment Online. In *International Conference on Web and Social Media (ICWSM)*. AAAI, 944–955.
- [43] Xinyue Shen, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns. In *USENIX Security Symposium (USENIX Security)*. USENIX.
- [44] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. *CoRR abs/1908.09203* (2019).
- [45] Charles Spearman. 1904. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology* (1904).
- [46] Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 22975–23005.
- [47] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. 2025. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 7184–7192.
- [48] Edward Tian and Alexander Cui. 2023. GPTZero: Towards Detection of AI-Generated Text Using Zero-Shot and Supervised Methods. <https://gptzero.me>.
- [49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR abs/2307.09288* (2023).
- [50] Urban Dictionary contributors. 2016. copypasta. <http://copypasta.urbanup.com/10468746>.
- [51] Urban Dictionary contributors. 2023. poojeet. <https://www.urbandictionary.com/define.php?term=poojeet>.
- [52] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* (2020).
- [53] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8692–8701.
- [54] Wikipedia. 2014. Pajeet. <https://en.wikipedia.org/wiki/Pajeet>.
- [55] Wikipedia. 2018. Pittsburgh synagogue shooting. https://en.wikipedia.org/wiki/Pittsburgh_synagogue_shooting.
- [56] Wikipedia. 2019. 2019 Balakot airstrike. https://en.wikipedia.org/wiki/2019_Balakot_airstrike.
- [57] Wikipedia. 2020. 2020–2021 China–India skirmishes. https://en.wikipedia.org/wiki/2020-2021_China-India_skirmishes.
- [58] Wikipedia. 2021. 2021 Israel–Palestine crisis. https://en.wikipedia.org/wiki/2021_Israel-Palestine_crisis.
- [59] Wikipedia. 2022. Ye antisemitism controversy. https://en.wikipedia.org/wiki/Views_of_Kanye_West.
- [60] Wikipedia. 2023. 2023 Israeli raid on Al-Shifa Hospital. https://en.wikipedia.org/wiki/Al-Shifa_Hospital_siege.
- [61] Wikipedia. 2023. Assassination of Hardeep Singh Nijjar. https://en.wikipedia.org/wiki/Hardeep_Singh_Nijjar.
- [62] Wikipedia. 2023. Jabalia refugee camp airstrikes (2023). [https://en.wikipedia.org/wiki/Attacks_on_Jabalia_refugee_camp_\(2023-present\)](https://en.wikipedia.org/wiki/Attacks_on_Jabalia_refugee_camp_(2023-present)).
- [63] Wikipedia. 2023. New York City synagogue tunnel incident. https://en.wikipedia.org/wiki/New_York_City_synagogue_tunnel_incident.
- [64] Wikipedia. 2023. October 7 attacks. https://en.wikipedia.org/wiki/October_7_attacks.
- [65] Wikipedia. 2024. Assassination of Hassan Nasrallah. https://en.wikipedia.org/wiki/Assassination_of_Hassan_Nasrallah.
- [66] Wikipedia. 2024. ChatGPT. <https://en.wikipedia.org/wiki/ChatGPT>.
- [67] Wikipedia. 2024. Columbia University pro-Palestinian campus protests during the Gaza war. https://en.wikipedia.org/wiki/Columbia_University_pro-Palestinian_campus_protests_during_the_Gaza_war.
- [68] Wikipedia. 2024. Political activities of Elon Musk. https://en.wikipedia.org/wiki/Political_activities_of_Elon_Musk.
- [69] Wikipedia. 2024. Self-immolation of Aaron Bushnell. https://en.wikipedia.org/wiki/Self-immolation_of_Aaron_Bushnell.
- [70] Wikipedia. 2024. Suicide of Atul Subhash. https://en.wikipedia.org/wiki/Suicide_of_Atul_Subhash.
- [71] Wikipedia. 2025. 2025 Bondi Beach shooting. https://en.wikipedia.org/wiki/2025_Bondi_Beach_shooting.
- [72] Wikipedia. 2025. 2025 Capital Jewish Museum shooting. https://en.wikipedia.org/wiki/2025_Capital_Jewish_Museum_shooting.
- [73] Wikipedia. 2025. 2025 India–Pakistan conflict. https://en.wikipedia.org/wiki/2025_India-Pakistan_conflict.
- [74] Wikipedia. 2025. 2025 Israel–Iran war. https://en.wikipedia.org/wiki/2025_Israel-Iran_war.
- [75] Wiktionary. 2024. street shitter. https://en.wiktionary.org/wiki/street_shitter.
- [76] Yahoo News Canada. 2025. 4chan "Operation Clog the Toilet" causes outrage. <https://ca.news.yahoo.com/4chan-operation-clog-toilet-causes-120110235.html>.
- [77] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. A Sanity Check for AI-generated Image Detection. In *International Conference on Learning Representations (ICLR)*.
- [78] W. J. Youden. 1950. Index for Rating Diagnostic Tests. *Cancer* (1950).
- [79] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM Internet Measurement Conference (IMC)*. ACM, 405–417.

A Open Science

To support reproducibility, we release code and derived artifacts and point to publicly accessible raw data sources:

Table 12: Correlation between Pangram AI detection rate and slur frequency, recomputed using multiple slurs ($N=49$ months).

Slurs counted	Anti-Indian		Anti-Semitic	
	ρ	p	ρ	p
Single slur (“pajeet” / “kike”)	0.57	<0.0001	0.21	0.15
All seven	0.57	<0.0001	0.20	0.18
All seven except “pajeet” / “kike”	0.54	<0.0001	0.17	0.25

- **Code.** All analysis scripts, including data collection, trend analysis, AI detection pipelines, toxicity comparison, and figure generation.
- **Derived data.** Aggregated monthly statistics (slur frequencies, AI detection rates, toxicity scores) and benchmark evaluation results.
- **Raw data.** The 4chan /pol/ data is publicly available through 4plebs.⁴

All code and derived data are available at our repository.⁵

B Ethical Considerations

This study analyzes publicly posted content on 4chan, an anonymous imageboard where all posts are anonymous by design. No personally identifiable information is collected or processed. Our AI detection pipeline is designed for aggregate trend analysis and should not be used to label or target individual users. We acknowledge the dual-use risk that our findings could inform adversaries about detection limitations; however, we believe the benefits of understanding AI-assisted hateful content outweigh this risk, as awareness is a prerequisite for effective countermeasures.

Researcher Well-Being. Our analysis pipeline is highly automated, limiting direct exposure to toxic content. Trend and peak

detection, AI-generated content detection, toxicity scoring, and image retrieval are all executed by scripts without manual inspection. Only two steps require human judgment. The first is peak classification, where researchers review posts to determine whether a peak is event-driven or discourse-driven. The second is CLIP threshold calibration, where a small number of images near the similarity boundary are inspected. To limit sustained exposure, this manual review is done in short, time-boxed sessions with regular breaks.

C Generative AI Usage

We used large language models (Claude and ChatGPT) as writing assistants during the preparation of this manuscript. Specifically, these tools were used for grammar correction, sentence-level rephrasing, and LaTeX formatting. In addition, Claude was used to generate initial code skeletons to support parts of the data processing and analysis pipeline. All AI-generated code was carefully reviewed, tested, and refined by the authors before use. All scientific content, including research questions, methodology design, data analysis, interpretation of results, and conclusions, was conceived and executed entirely by the authors. The authors take full responsibility for the correctness and integrity of all content.

D Robustness Beyond Single-Slur Measures

Table 10 measures slur frequency using the most frequent term for each target: “pajeet” for anti-Indian and “kike” for anti-Semitic hate. Table 12 repeats the analysis using all seven tracked slurs (Figure 4) and after excluding the most frequent slur for each target. The conclusions remain unchanged: the anti-Indian correlation remains strong and statistically significant even without “pajeet” ($\rho=0.54$, $p<0.0001$), whereas the anti-Semitic correlation remains non-significant.

⁴<https://archive.org/details/@fourplebs>.

⁵<https://github.com/TrustAIRLab/GenAI-Hate>.