The Rise of ‘AI Slop!’ Accusations Is Becoming a New Form of Gatekeeping

Originally published at https://www.unite.ai/the-rise-of-ai-slop-accusations-is-becoming-a-new-form-of-gatekeeping/

Calling something ‘AI slop’ has become the internet’s new witch-hunt, with Reddit and Hacker News users increasingly accusing fellow commenters of being robots, even when there’s no evidence of this.

A new study from Norway and the UAE has found that accusations against supposed ‘AI slop’ from other commenters rose sharply across Reddit and Hacker News between 2023 and 2026, even when the comment showed no evidence of being AI-generated.

Results from the authors’ analysis of 25 million comments suggest that such accusations are increasingly functioning as an emergent form of *social gatekeeping*, rather than as a way to identify AI.

The researchers also found that technically-minded communities adopted the ‘accusation culture’ earlier than other groups, with the pattern later spreading across wider areas of Reddit.

This apparent rise in accusations around ‘AI slop’ does not appear to be part of a broader or more general increase in online hostility: older invective forms such as *‘shill’*, *‘sockpuppet’*, and *‘troll’* remained relatively stable during the same period, suggesting that suspicion of AI emerged as a novel form of social boundary-policing, rather than a continuation or extension of earlier internet feuds.

The paper states:

> *‘We analyzed 25 million comments from Hacker News and Reddit (2023-2026), combining LLM judgment on 7,500 sampled accusations of AI use, sentiment trajectories, speech-act coding of 300 confirmed accusations of AI use, and a matched-control test of accused versus non-accused parent comments.*
>
> *‘We found that the pejorative-label share of accusations rose more than tenfold on both platforms while a placebo vocabulary of pre-2022 inauthenticity terms (“shill”, “astroturf”) did not.*
>
> *‘This shift reflected a fast-growing trend of branding any suspicious or seemingly inauthentic prose as “AI slop”.*
>
> *‘The slop frame now constitutes 94 percent of pejorative mentions, with the dominant comments shifting in tone from mockery toward gatekeeping and structural protest.’*

The study raises the broader question as to whether people can really spot AI writing, since fluent prose – formerly treated as evidence of effort, expertise, or genuine engagement – is now an abundant and increasingly devalued commodity.

It’s interesting to note that the new work concentrates on Hacker News, which is vigilantly policed against AI-generated comments, and on Reddit, whose constant flow of human-based discourse is now highly-prized for AI developers and companies, as well as becoming a new prime target for SEO spammers looking to invade LLM-based web rankings by proxy.

The researchers believe that their findings accord with the growing public understanding that prior sources of truth could be devalued as the use of AI spreads. The new paper discusses real people accused of being AI entities, either through genuine error, stylistic conflation, or malice (i.e., the accuser knows their opponent is human, but wishes to shut them down); but predicts other types of communication being similarly tarnished:

> *'Our results here would predict that similar AI-use accusations will form for image authentication, voice authentication, and code authorship among others, with the core intent of the lay accusation being gatekeeping rather than empirically accurate detection of AI use.*
>
> *'This will become increasingly problematic as AI in those areas reduces even the empirically detectable tip-offs that experts can find.*
>
> *'This could have the effect of increasing the role of experts in verifying AI vs. non-AI content; or it could greatly reduce trust in any type of medium that can be plausibly generated by AI.'*

The new paper* is titled *"That's AI Slop, You Bot!" Studying Accusations, Evidence, and Credibility in Online Discourse Towards LLM-Generated Comments*, and comes from two reviewers across the University of Oslo and the American University of Sharjah.

Method

The dataset developed for the new study comprised all public comments posted to Hacker News and 18 selected Reddit communities between January 2023 and May 2026.

Around 25 million comments were curated, with 12 million from Hacker News, and 13 million from Reddit. Reddit data was obtained from the Arctic Shift archive through its public JSON API, while Hacker News comments were collected from the Algolia Hacker News search archive.

To avoid focusing on a single type of community, the Reddit sample was divided across AI-focused forums including *r/aiwars*, *r/ArtistHate*, *r/ChatGPT*, *r/OpenAI*, *r/MachineLearning*, *r/LocalLLaMA* and *r/singularity*; creative communities comprising *r/Art*, *r/writing* and *r/books*; the general-interest forums *r/AskReddit*, *r/news*, *r/changemyview*, *r/explainlikeimfive*, *r/AskHistorians* and *r/science*; and the technology-oriented and academic communities *r/programming* and *r/AskAcademia*.

Sampling rates were kept consistent across time, helping to ensure that changes in accusation rates reflected shifts in community behavior, rather than changes in data collection.

Five Levels of AI-Shaming

Candidate comments were identified using a 137-pattern search lexicon organized into five named tiers: Tier 1 (*'Direct'*) captured explicit accusations such as *'ChatGPT wrote this'*, *'Is this AI-generated?'*, and *'OP is a bot'*.

Tier 2 (*'Pejorative'*) covered labels such as *'AI slop'*, *'GPT garbage'*, *'ML drivel'*, and *'robo-writing'*. Tier 3 (*'Style'*) dealt with supposed stylistic tells, including em-dash mentions, the 'delve' callout, tricolon references, and broader claims about a 'classic AI signature'.

Tier 4 (*'Mocking'*) captured parody and imitation based on familiar AI-assistant phrases such as *'fellow humans'*, *'in the rapidly evolving landscape'*, and *'rich tapestry'*. Tier 5 (*'Indirect'*) featured less explicit suspicions, with comments described as something that *'smells like AI'*, *'reads like ChatGPT'*, or resembles the *'uncanny valley of writing'*.

To reduce false positives, common phrases such as *'worth noting'*, *'it's important to note'*, and *'is this a human'* were only counted when an AI-related term appeared nearby. Because these search patterns could not reliably distinguish accusations from ordinary discussion, two validation passes were subsequently carried out with Claude Opus 4.7.

A Reddit sample of 5,000 comments and a Hacker News sample of 2,500 comments were drawn from the candidate pool, balanced across time periods and accusation categories.

Each comment was then classified into one of five outcome groups: *Real*, covering genuine accusations of AI use; *Disclosure*, covering comments that acknowledged AI authorship; *Neutral-Ref*, covering non-accusatory references to AI; *FP*, covering regex false positives; and *Ambiguous*, covering cases where the available context did not permit a confident judgment.

The researchers also examined how accusations changed over time, tracking the rise of the newer *'AI slop'* framing against older insults such as *'drivel'*, *'garbage'*, *'trash'*, '*vomit'*, *'sludge'*, *'mush'*, *'gunk'*, *'junk'*, *'crap'*, *'word salad'*, and *'nonsense'*.

Delimiting the Trends

Sentiment trends were measured using Valence Aware Dictionary and sEntiment Reasoner (VADER), while a separate sample of 300 Reddit threads containing LLM-validated *Real* accusations was coded according to the social role being performed. These were classified as *Sneer* (dismissive mockery); *Dismiss* (straight rejection); *Mockery* (imitation/parody); *Gatekeep* ('rule-enforcement'); or *Structural Protest* (general disapproval of AI), allowing shifts in the character of AI accusations to be tracked over time.

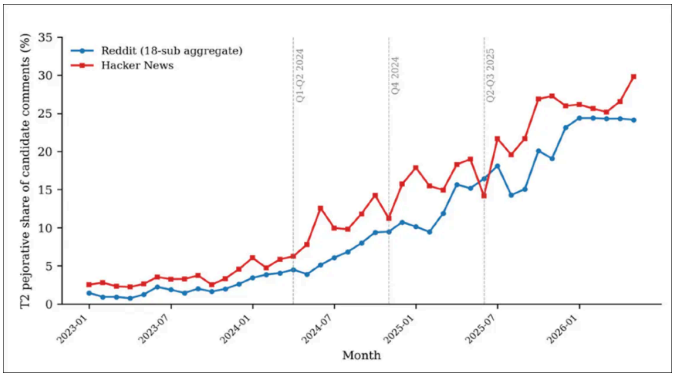
A separate 'placebo' test was designed to determine whether rising AI accusations might simply reflect a broader increase in suspicion online, wherein the same dataset was searched for older pre-ChatGPT inauthenticity terms such as *'shill'*, *'astroturf'*, *'sockpuppet'*, *'paid shill'*, *'fake account'*, *'corporate shill'*, *'talking points'*, and *'payola'*.

A final set of tests examined whether the traits that separate AI-generated writing from human writing are the same traits that cause human-written comments to be accused of being AI, through the examination of six linguistic markers: *article density*; *contraction rate*; *formal-register adverb frequency*; *preposition density*; *sentence-length variance*; and *mean token length*. Comparisons were made between *Disclosure* and *Real* comments using Mann-Whitney U tests.

Parent comments associated with 800 LLM-validated Real Reddit accusations were retrieved, with 421 cases retained where the parent was itself a comment rather than a top-level post. These were matched against 2,048 non-accused comments drawn from the same subreddit and month. Logistic regression was then used to test whether the linguistic markers that distinguish AI-generated text from human writing also predict which human-written comments attract accusations of AI use.

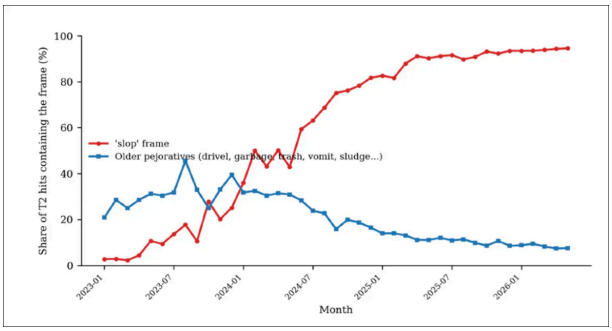
Results

The study recorded a large increase in AI accusations across Reddit and Hacker News between 2023 and 2026. Most of this growth was concentrated in the use of pejorative labels:



Cross-platform growth in pejorative AI accusations on Reddit and Hacker News between January 2023 and May 2026. Tier 2 ('Pejorative') accusations rose from low single digits to roughly one quarter of candidate accusations on both platforms. Three acceleration periods are visible during 2024 and 2025, after which growth leveled off. Hacker News remained above Reddit for most of the study period, but both converged on similar levels by 2026. Source

By 2026, 'AI slop' accounted for 94% of pejorative AI accusations identified in the dataset, replacing earlier terms such as *'GPT garbage'*, *'ML drivel'*, and *'robo-writing'*. According to the paper, the share of pejorative AI accusations increased by more than tenfold on both platforms during the study period:



Rise of the 'AI slop' label relative to older pejorative AI accusations between 2023 and 2026. While terms such as 'drivel', 'garbage', 'trash', 'vomit', 'sludge', 'mush', 'gunk', 'junk', 'crap', 'word salad', and 'nonsense' initially dominated pejorative accusations, their share declined steadily as 'AI slop' became the overwhelmingly preferred label. By 2026, the 'slop' frame accounted for roughly 94% of pejorative AI accusations, indicating a consolidation of accusation language around a single term.

A separate comparison was carried out using older inauthenticity terms comprising *'shill'*, *'astroturf'*, *'sockpuppet'*, *'paid shill'*, *'fake account'*, *'corporate shill'*, *'talking points'*, and *'payola'*. Unlike AI-focused accusations, these terms did not exhibit a comparable increase.

Variation was also observed across communities, with earlier growth recorded in AI-focused and technology-oriented forums – with similar patterns later appearing in other parts of Reddit and Hacker News.

Changes were observed not only in the frequency of accusations but also in their classification. Coding of 300 validated Reddit accusations found shifts in the relative prevalence of *Sneer*, *Dismiss*, *Mockery*, *Gatekeep*, and *Structural Protest*. According to the paper, *Gatekeep* and *Structural Protest* became more common over time, while *Sneer* and *Mockery* became less common.

Conclusion

The apparent epidemic of casual AI-shaming in comments sections clearly needs its own iteration of Godwin's Law; based on the events and trends in social and political commentary of recent years, it would make sense if AI bots were to become the *most likely to accuse other commenters of being a bot*; however, this might tend to stifle all commentary on the matter.

* *Please be aware that this paper is not a friendly read, and is aimed, in tone and lexicon, at the authors' academic peers.*

First published Friday, June 12, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More