Is This AI? Longitudinal Analysis of Strategies Used for AI Detection on Two Subreddits

Christina Yeung[1], **Galen Weld**[1], **Jaron Mink**[2], **Franziska Roesner**[1]
[1]University of Washington
[2]Arizona State University
cyeung3@cs.washington.edu, gweld@cs.washington.edu, jaron.mink@asu.edu, franzi@cs.washington.edu

Abstract

As AI-generated content (e.g., "slop") becomes more prevalent online, people are developing strategies to attempt to identify it (or, conversely, to gain confidence that something is not AI-generated). What strategies are people using, and how are they changing over time as generative AI models themselves change? In this work, we catalog and analyze 2 years and 8 months of the AI detection strategies discussed by users of two popular Reddit communities (r/isthisAI and r/RealOrAI) that use the wisdom of crowds to identify AI-generated media. Through a mixed-method analysis of 13,098 posts and 222,060 comments within these communities, we catalog and analyze the prevalence of 12 AI-detection strategies, including examining fine-grained physical details, recognizing trends in AI-created content, and the assumptions people make about what models are capable of producing. Furthermore, we find that these strategies and mental models shift over time in accordance with changing AI capabilities and in response to online social trends. By systematically cataloging users' AI detection strategies, we lay the groundwork for user-facing guidance and future research.

Introduction

The last few years have seen tremendous growth in generative AI in terms of model capabilities, the accessibility of these tools to end-users, and the proliferation of AI-generated content online, including text, images, audio, and videos. Though these capabilities and tools are exciting for some use cases and some users, they also raise myriad concerns, including the potential for "deep fakes" to spread political or other disinformation (Naitali et al. 2023), interpersonal abuse (e.g., non-consensual generated intimate imagery) (Gibson et al. 2025), and the proliferation of low-quality content on the web and social media (often called "AI slop") (Hoffman 2024; Shaib et al. 2025).

As a partial mitigation to some of these concerns, it may be helpful for users to be able to distinguish genuine content (e.g., human-generated text or unaltered photos/videos) from AI-generated content online. For example, correctly identifying AI-generated content can help users avoid political misinformation or scams; on the other hand, believing something AI-generated is real may contribute to the spread of misinformation. However, as generative AI techniques become increasingly sophisticated, it has also become increasingly difficult for people to tell the difference just by looking: for instance, an AI-generated video clip of bunnies jumping on a trampoline went viral in 2025, fooling people into thinking that it was authentic (Thomas 2025).

There are several approaches to help remedy the difficulty of identifying AI-generated content, though each with their limitations. AI watermarking schemes are promising, yet are currently easily bypassed by simply using a tool that does not watermark its output (Christ, Gunn, and Zamir 2024; Zhao et al. 2025; Liu et al. 2024; Dathathri et al. 2024; Hussain et al. 2021). Furthermore, automated AI detection tools are subject to an ever-escalating 'arms race' between people generating AI content and those seeking to detect it (Weber-Wulff et al. 2023; Gotoman et al. 2025; Liang et al. 2023; Layton et al. 2026). On the UX side, some social media platforms explicitly label AI-generated content, but these labels are only as good as the method for determining whether something is AI-generated in the first place.

Thus, users and online communities are increasingly taking matters into their own hands: identifying AI-generated content via comments, creating and following social media accounts that aim to educate users in identifying AI-generated content (JeremyFindsAI 2026), and turning to social media groups dedicated to AI content identification. Such groups have sprung up on (for example) Facebook and Reddit, and provide hundreds of thousands of discussions of potentially AI-generated content, offering insights into how people perceive and attempt to detect such content. Since people act on their perceptions of the content they see, it is important to understand how people form those perceptions.

However, to the best of our knowledge, there has been no systematic study of what rationales or strategies are employed by internet users to identify AI-generated content in the wild, which limits efforts towards user education and platform design for content labeling. In this work, we study two communities on Reddit to investigate the strategies used to assess content — and we do so longitudinally over the lifetime of these communities (2 years and 8 months). We investigate the following research questions:

RQ1 What strategies are used to identify AI-generated content within two prominent subreddits (r/RealOrAI and r/isthisAI)?

RQ2 How do strategies change or vary in different contexts, e.g., over time or across different types of content?

To answer these questions, we collect 13,098 posts and 222,060 comments from two large subreddits, r/RealOrAI (established in 2022, with 1,700 weekly contributions) and r/isthisAI (established in May 2023, with 22,000 weekly contributions), and create a taxonomy of strategies people use to detect AI-generated content by qualitatively analyzing the comments using a combination of human and LLM annotations. We identify and describe the strategies employed by social media users in the wild, and characterize the prevalence of different strategies over time and across different types of content; we do not directly evaluate the effectiveness of the strategies themselves.

Our contributions include the taxonomy of strategies, as well as qualitative and quantitative observations gleaned from our data. For instance, while we find that people most often use physical cues, such as hands or teeth to detect AI-generated content, we also find evidence to suggest that people are increasingly moving away from relying on fine-grained details, and instead using mental models developed over time about model capabilities and AI trends.

Our work provides systematic documentation of strategies and rationales used by members of two prominent online subcommunities whose mission is to help others identify AI-generated content (or give confidence that something was not AI-generated). By systematically cataloging users' AI-detection strategies, we lay the groundwork for user-facing guidance and future research. We make our dataset available at https://agent-security.cs.washington.edu/is-this-ai.html.

Related Work

Perceptions of AI-Generated Content. Researchers have documented that people's perceptions of content generated using AI is often more negative than their view of content created by other humans. For instance, academics at the University of Florida found that people reported disliking stories they thought were written by AI, even if other people actually wrote them (Chu and Liu 2024). Another study found that people credited artists who used generative AI in competitions with creativity and effort, but did not believe that it required skill (Lima, Grgić-Hlača, and Redmiles 2025). And, a recent Pew report found that most people report being concerned about the growing use of AI in daily life (Kennedy et al. 2025). Amongst other findings, the Pew report highlighted people's concerns that they would not be able to distinguish what was, and was not generated by AI. And these concerns seem well founded: Microsoft reported that people were not skilled at distinguishing AI generated images from "real" images when comparing them side-by-side, particularly when it came to evaluating natural and urban landscapes (Roca et al. 2025), or Cooke et al. who find that people's ability to detect AI generated content is no better than a coin toss (Cooke et al. 2025). While our work does not directly investigate people's perceptions of synthetic, AI generated content to "organic" content, we also observed comments that supported both the sentiment that synthetic content was inferior, as well as users who expressed concerns about their ability to distinguish AI-generated content as models progressed in capabilities.

Automated Detection of AI-Generated Content. As AI use has grown so have automated tools to detect AI-generated content. This is particularly true in the context of detecting AI use in academic settings: several startups market themselves as plagiarism checkers, including GPTZero, Originality.ai, Copyleaks, and DetectGPT.

However, there is also skepticism about the accuracy and robustness of AI detection tools. For instance, researchers have found that there are easy ways to evade some tools, from using a single space (token mutation cases) (Cai and Cui 2023), paraphrasing generated content (Krishna et al. 2023; Sadasivan et al. 2025), and using LLMs themselves through prompts to intentionally evade detectors including prompting imperfections in the output created by generative AI (Lu et al. 2024). Detection tools have been found to be unreliable (Layton et al. 2026; Ernst, Rupp, and Simbeck 2025), including exhibiting biases or systematic errors. For example, content from non-native English speakers is more often falsely identified as AI-generated than content from native English speakers (Weber-Wulff et al. 2023; Gotoman et al. 2025; Liang et al. 2023). Further, detection tool performance is brittle to model updates, such as GPT 3.5 to 4 (Elkhatat, Elsaid, and Almeer 2023). Researchers have also found that some methods proposed to make generated content detectable, such as watermarks, can easily be removed using generative AI (Zhao et al. 2024).

In our results, we find limited use of AI detection tools (in part due to subreddit rules), and find that users also rely on a wide range of other, more ad hoc strategies.

Human Detection of AI-Generated Content. There has also been work that examines people's strategies and abilities to distinguish AI-generated content in different contexts, such as on LinkedIn profile pictures (Mink et al. 2022, 2024), images of faces (Bray, Johnson, and Kleinberg 2023), AI generated videos (Fu et al. 2025), audio clips (Mai et al. 2023; Diel et al. 2024; Müller, Pizzi, and Williams 2022), deepfakes (Farid 2022). Our work similarly examines people's mental models and reasoning they use to distinguish AI generated content, but does so from the perspective of posts and comments from two subreddit communities, which means that the content that people discuss may overlap with the types of media discussed in prior work. However, because Reddit users can post about anything, our work provides a broader look into a variety of content.

Finally, users may be directly informed that content is AI-generated via platform labels on content; recent work has found that people tend to over-rely on these signals (Höltervennhoff et al. 2026).

Methods

To understand strategies people use to identify AI-generated content at scale and over time, we performed a large-scale mixed-method analysis of 222,060 crowd-sourced, free-text responses on whether 13,098 pieces of media were, or were not, made with AI. We discuss the ethical considerations in the Ethics and Adverse Impacts Appendix.

Subreddits of Focus. To gain a high-resolution understanding of how human identification of AI occurs today and has changed in recent years, we leverage public posts and responses made in two subreddits: 'r/RealOrAI', and 'r/isthisAI'. Both of these subreddits exist to provide crowd-sourced answers to whether posted digital media (often photos, videos, or even audio clips), is or is not made with AI. While other subreddits exist (e.g., r/AIorFake (Reddit 2026a), r/AI_or_Real (Reddit 2026b)), we focus on r/isthisAI and r/RealOrAI as they are substantially larger and more active, with 29,000 and 1,700 average weekly contributions (including posts and comments), respectively. We collect our data from both subreddits using ArcticShift (Heitmann 2024), a widely available tool that is intended to make Reddit data available to researchers.

Data Collection and Filtering. We collect every post and comment from r/isthisAI and r/RealOrAI from the communities' creation dates (August 2022 and May 2023, respectively) through March 2026, encompassing 2.8 years of data.

Posts. To ensure our analyzed posts center on human responses to whether the posted media is or is not AI, we filter our collected posts that: (1) Focus on soliciting crowd wisdom; while all posts from r/isthisAI are requests for crowd-sourced human strategies for determining AI, only the posts labeled "[HELP]" in r/RealOrAI have this focus, and thus are the only sub-set of posts we collect,[1] (2) Are valid; in particular, we remove all posts that have been deleted by their the author or a community moderator, (3) Are relevant; lastly, we go through all posts, and any that do not focus on crowd-sourced responses to media are labeled as "irrelevant" and removed from the dataset and subsequent analysis. In total, we analyze 13,098 posts across the two communities (r/RealOrAI: 36.6%, r/isthisAI: 63.4%).

Comments. From our filtered posts, we then collect all user-submitted comments and specifically analyze the subset of them that are: (1) Direct comments to the post; while comments made generally on the post may be responding to other users' comments, often on tangentially related topics, direct comments are often in response to the post-purposes and thus are a crowd-sourced response to whether the media is or is not AI. (2) Are valid; we remove all comments that have been deleted, removed, or authored by users who have deleted accounts. (3) Are relevant; like posts, we label and remove comments coded as "irrelevant". In total, we analyze 222,060 comments (r/RealOrAI: 46.6%, r/isthisAI: 53.4%).

Qualitative Analysis: Human Coding. To characterize the strategies humans use to detect AI, we performed a hybrid qualitative coding methodology (Azungah 2018) to taxonomize these strategies and measure the prevalence across Reddit. To develop an initial codebook, a single coder began with specific codes selected from Fu et al. (Fu et al. 2025) and the SIFT method (Caulfield 2019), a digital guide designed to help people establish truthfulness in what they observe; using a randomly selected set of 500, the single coder than inductively added new codes, such as CONTENT_PLAUSIBILITY, AI_STYLE, and MODEL_CAPABILITIES, among others.

Once the initial codebook reached conceptual saturation and no new codes emerged, it was given to a second coder to calculate inter-rater reliability. One hundred randomly selected comments were independently coded by two coders, and the inter-rater reliability (IRR) of each sub-code was then calculated using Cohen's κ (Cohen 1960). After coding, the coders met to calculate the codebook's IRR, and resolve disagreements; if any sub-code did not achieve a high inter-rater reliability ($\kappa > 0.7$), this process was repeated. After three rounds, the IRR for all subcodes was achieved after 300 comments in total. We present the final codebook and IRR values in the full extended version of the paper (Yeung et al. 2026). Once this codebook was established and verified for reliability, we manually coded 2,000 randomly selected comments and used the resulting human-coded set as ground truth to design and verify our LLM-assisted coding methodology.

Qualitative Analysis: Automated Labeling. To label all 222,060 posts and comments based on our taxonomy, we developed an automated labeling method and validated it on held-out data. We used a zero-shot classification pipeline built around Claude Sonnet 4.6 (Anthropic 2026), structuring each of our 13 codes as an independent classification task. Using 1,000 manually-coded and randomly sampled comments to serve as a development set, we iteratively refined our prompt until we achieved high inter-rater reliability between the automated system and the ground-truth human labels, with Cohen's $\kappa > 0.7$ (Kraemer et al. 2012). To validate performance, we then ran our system on an additional *holdout* set of 1,000 human-annotated comments that the automated system had not seen during prompt refinement. Within this set, we found that every subcode still had substantial agreement ($\kappa > 0.84$ for every subcode). We then use this AI system to code our full dataset of posts and comments (N=222,060), which we analyze and present in our results. We provide the final prompt, along with additional details on the process of creating it, in the extended version of the paper (Yeung et al. 2026).

Quantitative Analysis: Frequent Words. To further characterize strategies, we explore the important words that commonly appear. We do this using tf-idf (Salton and Buckley 1988), treating the comments within each sub-code as a "document". Prior to running tf-idf, we do not stem words (i.e., reduce to roots), and we remove non a-z, 0-9 characters (e.g., removing punctuation). We ignore terms that appear in more than 90% of documents, and identify the top 10 most unique words for each sub-code.

Quantitative Analysis: Temporal. Given our full set of labeled comments, we also conducted several quantitative analyses towards answering RQ2, i.e., comparing the frequency of different strategies in comments over time.

Our dataset covers about 33 months, from July 2023 through March 2026. To observe temporal trends, we con-

[1] Other tags, such as "[GUESS]" ask users to respond to a user-made challenge, but are not requests for helping determine the authenticity of content by the requester. We also manually identified 136 posts on r/RealOrAI that *would* have been "[HELP]" posts (i.e., poster asked for help and did not provide an answer) but appeared before this guideline was enforced in late 2024.

Category or Sub-code	Count	%	Description
Perceptions of World			
PHYSICAL_DETAILS	119,390	44.5%	Examines fine-grained details in media
CONTENT_PLAUSIBILITY	66,009	16.8%	Compares to personal experiences and real world
Perceptions of AI			
AI_STYLE	36,039	9.0%	Assumptions about characteristics of AI content (e.g., texture)
MODEL_CAPABILITIES	15,597	3.0%	Assumptions about what models are and are not able to do
MODEL_PROMPT_SPECULATION	7,653	1.5%	Hypotheses about what models or prompts were used
External Signals			
TRIANGULATING_INFO	21,194	5.3%	Identifies original poster or searches for corroborating evidence
MEDIA_AGE	7,191	2.1%	Compares date of content with model capabilities at that time
TOOL_SIGNALS	4,445	1.4%	Use of tools including reverse image search, watermarking, etc.
OTHER_EXTERNAL	1,385	0.2%	Use of other external sources
Perceived Motivations for AIG Content			
AI_TRENDS	6,957	1.8%	Uses cues from current events, pop culture, other viral content
MOTIVATION_FOCUSED	6,818	1.6%	Speculates about reasons others might create/spread AI content
Intuition	28,195	12.7%	Short, decisive comments without support or reasons

Table 1: Taxonomy of AI detection strategies in our dataset, including the number of comments in which each strategy appeared, and the percentage this represent of all comments (N=222,060). Note that a single comment may describe multiple strategies.

sider both monthly rolling averages as well as three 11-month segments of our data. These are: *Period 1:* July 7, 2023 – May 31, 2024; *Period 2:* June 1, 2024 – April 30, 2025; and *Period 3:* May 1, 2025 – March 31, 2026.

In both analyses, we compare not the raw number of times a strategy appears (since overall comment volume varies over time) but normalize to the total number of strategies mentioned in each period. That is, we calculate the relative proportion of that strategy among all strategies mentioned during that time period. We normalize by the number of strategies mentioned, rather than the number of comments, since a single comment can contain multiple strategies.

We complement visualizations with statistical tests. We first run χ^2 tests to compare the distribution of strategies observed between periods, correcting for multiple comparisons with Bonferroni. To more closely examine how individual strategy proportions change over time, we also run post-hoc tests, again with Bonferroni corrections.

Limitations. Our work has limitations similar to prior work that also analyzes general perceptions via Reddit (Bouma-Sims et al. 2024; Oak and Shafiq 2025; Bouma-Sims, Lanyon, and Cranor 2025; Wei et al. 2024; Nath et al. 2026). *First*, while we analyze users' AI detection strategies across two of the most popular crowd-sourced AI-detection communities, it is unclear if the perceptions and strategies here generalize to the general population, who may be less aware, less interested, and interact less with shared opinions for how to detect AI effectively. To mitigate, we do our best to contextualize our findings for these communities, while also arguing that the trends within this active group of users continue to provide valuable signals about how beliefs and perceptions of AI have changed over time. *Second*, while we leverage ArcticShift (Heitmann 2024) to acquire data, we are also beholden to their data scraping methods and preparation; in particular, this may not reflect the current version of Reddit in which data is later deleted, and also presents comments as tied to the date of the post's creation. To mitigate this, we do our best to remove data and comments that have since been deleted on the live site, and while we cannot track comments to a date beyond the post, we also argue that as active subreddits with thousands of new posts a day, the vast majority of comments likely occur within a few days of the post's creation. *Third*, while our analysis provides insights into how posted strategies have changed, we ultimately do not claim to identify the root causes of these changes. While we provide hypotheses based on users' self-reported reasoning in posts and technological changes around that time, we ultimately leave this investigation for future work to confirm. *Fourth*, We acknowledge that there are many ways to analyze this dataset, such as how media-specific or user-specific characteristics affect posted comments. We encourage future researchers to leverage our dataset or approach to explore these questions from different perspectives. *Fifth*, it is reasonable to question whether our AI-assisted methodology may result in biased data that differs from human coding. To the best of our ability, we mitigate this by comparing our AI-assisted coding to a ground truth of human-coded data made after our codebook's IRR was established. Thus, we argue that our AI-assisted codings are, by definition, as empirically reliable as any other IRR-verified coding methodology.

Results

Overview of Dataset. Our dataset comprises 222,060 comments across r/RealOrAI (103,480, 46.6%) and r/isthisAI (118,580, 53.4%) between July 2023 to March 2026. Much of the data comes from the end of 2025 and beginning of 2026 as both subreddits grew in activity. (Additional data on post and comment counts over time is in the extended version of the paper (Yeung et al. 2026).)

RQ1: Strategies to Identify AI-Gen. Content

As shown in Table 1, we find a range of strategies that are used by users to determine whether a piece of media is made with AI. While we find these strategies leverage one of three mental models (Mink et al. 2024) for how AI media looks (**Perceptions of AI**) or what the real world *doesn't* look like (**Perceptions of World**), we also find heavy use of **External Signals**, and unspoken **Intuition**. We now describe each of these individual codes by their general strategies.

Perceptions of World

To determine whether media was or was not AI, it was often contrasted with what people experience in the real world, and how they differ, or align. We find that people often focus on what is (not) physically possible, and whether the phenomena captured appear (un)known or (un)common.

Physical Details. 44.5% of comments examined detailed features within content, whether those were 'aligned' with physical reality or physically impossible. The important words from tf-idf (see extended version of the paper (Yeung et al. 2026)) include specific objects being scrutinized (e.g., *handles*, *tiles*) as well as physical features (e.g., *pupils, elbow* and *index* (finger)).

As people and animals were often the subject of generation, many comments focused on abnormalities within them, such as their hands, fingers, eyes, teeth, whiskers, paws, and fur. For instance, in a post where the original poster was looking for help analyzing the prints on two T-shirts they bought, one comment highlighted how the hands depicted in the image were incorrect, saying: *"...they are both AI... the first one is the hand holding the cat, a finger going the wrong way"* (User 71185). People also examine the texture and pattern of elements in the content. They might bring up how someone's skin is too shiny, how the texture of clothing looks incorrect, or how the content appears overall distorted or 'melty'. For instance, when evaluating art, one person said: "*the eyes look melty and classically AI"* (User 9798).

The disruption of universal physical laws, such as inconsistent lighting, shadows, reflections, or focus, was also a clear indicator used by people. For instance, considering whether promotional content from an electronics manufacturer used AI, part of someone's comment focused on the shadows that ought to have been cast, saying: *"Also, the light to the right side of the desk should cause a double shadow on the mini PC. The light from the window seems to be the only light actually casting shadows. The light to the right of the desk is there but it doesn't look like anything is interacting with that light"* (User 52940).

For videos, consistency of these factors over time was a common cue. Odd motion, or how objects change over time, often caused suspicion: *"The faces are so so wrong looking, they don't move right, this is definitely AI"* (User 67645).

While these details are often used to justify that something is AI-generated, a few were used to rationalize authentic content. For instance, the realistic motion and the consistent lettering in the video clip of a child on rollerskates was used by one person to justify that the media didn't involve AI: *"[I am] Leaning real. [The] Letters on the shops don't change or morph"* (User 9632).

Content Plausibility. 16.8% of posted comments broadly assess whether the content aligns with their expectations of how people, animals, and the world behave. Unlike PHYSICAL_DETAILS which tends to focus on minor, but clear inconsistencies reminiscent of AI creation, comments here focus on general behavior trends that are common or uncommon. For instance, when discussing a video clip showing baby and a cat interacting, one user comments: *"Babies do not nearly have that much coordination to both be standing up, talking so well and holding up that large cat. It is obvioulsy [sic] AI"* (User 12140). Similarly, another person comments on a clip of horses, saying, *"I don't see anything screaming fake. The horses are looking at/reacting to each other realistically. "* (User 4581).

Perceptions of AI

Often, people determine what is AI based on their preconceived notions of what AI. In particular, users often focused on what AI content tends to look like, what they believe the limits of AI capabilities are, and what content tends to be generated from common prompts.

AI Style. 9% of posted comments consider whether the content looks like or includes tell-tale signs of common AI-generated outputs. For instance, one user began their comment by saying *"Several classic hallmarks of AI image generation are evident..."* (User 33021). These tells vary by the media type being evaluated. When considering text, people often noted that the presence of em dashes or the use of particular sentence structures (e.g., examples in groups of three, generic or overly ornate but meaningless phrases) was indicative of AI-generated content. Comments on images often consider specific font choices in images, and while comments on audio often look for 'generic' speech. In video, this is often seen in overly smooth motion, emotionless eyes, and specific events such as yelling in the first few seconds or the movement of liquids, such as water, which are common in AI-generated videos. For instance: "*...If you've seen other AI videos, you know, that liquid motion is one of the hardest things for it to emulate naturally*" (User 56774).

Table 2 lists the words commonly (uniquely) used to describe AI style, including words describing the texture of content (e.g., *glossy* and *sheen*). There are also specific markers that people associate with AI content (e.g., *dash* as in em-dash), and specific color palettes (e.g., *yellowish*, *hue*). For instance: "*It's not just misspelling but the em dash, and the kerning. A human didn't write that*" (User 74675).

Some people noted limitations of this approach, observing that because models are trained on data from existing artists, their outputs could mimic the styles of those artists — making the original artists' work appear, now, to be (incorrectly) stylistically AI. For instance, someone said, " *...you're also making it harder for real artists. Some of you are convincing people that *actual* artistic quirks, mistakes, and techniques are 'obvious' AI. AI was partially trained on real, stolen, art"* (User 63442).

Model Capabilities. 3% of posted comments leverage the

AI STYLE	AI TRENDS	MEDIA AGE	MODEL / PROMPT SPECULATION	MOTIVATION FOCUSED	TOOL SIGNALS
glossy	popping	preai	diffusion	butchering	synthid*
soulless	flooded	2016	dalle	crypto	wwwtorfai
sheen	doorbell	spaghetti	grok	ragebait	tineye
dashes	spaghetti	covid	generates	waters	isthisaicom
outlines	chihuahua	2014	chatgpts	scammy	sightengine
flour	trends	january	typed	victim	aidetective*
yellowish	sleeping	dated	flux	goal	watermarked
cartoony	trampoline	predate	input	gain	cybyapi
brushes	porch	2013	prompter	sympathy	assigned
hue	react	october	brushes	motivation	analyzed

Table 2: Top 10 unique keywords in specific subcodes (resulting from tf-idf). Note that punctuation has been removed: we shortened some urls in TOOL SIGNALS and represent this with an asterisk. Words are ordered in decreasing tf-idf value.

users' perception of what AI systems can, and cannot, do. For instance, people might assume that AI is not capable of creating realistic water movement, and use that assumption as the basis for believing that all beachside images must therefore be authentic. Another common example was assumptions around models' (un)abilities to produce legible, well-kerned text in images. Other examples connect with the details that users identify under the PHYSICAL_DETAILS and AI_STYLE sub-codes: e.g., models' inability to keep moving objects consistent, to represent physics realistically, or to use certain art styles in appropriate contexts. For instance, one user opining on a video wrote, *"AI still doesn't seem to understand eye movements somehow"* (User 18658).

People also considered how models could outperform humans. For example, after identifying the original account responsible for YouTube videos, people used the fact that this account frequently posted long (over 30 minute) videos to determine that it must be AI-generated, as only AI could produce that volume of content so quickly.

Interestingly, we find that these beliefs are not only related to inherent model capabilities but also to the quirks of particular AI systems and side channels they may have. For instance, AI systems such as Sora 2 allowed users to generate clips of up to 15 seconds of high-quality, realistic video clips, even if they did not pay for the service (Peckham 2025). As these systems were both commonly used and posed a limit to what lay-AI creators could generate, people often used them as indicators to question whether AI was used, as in this case with User 112525 who said *"Well considering it's longer than 15 seconds it probably not AI"*, as to confirm AI's presence: *"[the] Video is also exactly 10 seconds long.. not looking good"* (User 74212).

Model and Prompt Speculation. Another strategy people use to support their assessment that something is AI-generated is to speculate about the models or prompts that were used to generate the content in question (the MODEL_PROMPT_SPECULATION sub-code). For example, people often referred to specific AI companies when identifying elements in content that tipped them off to believe it was AI-generated: e.g., "ChatGPT-style" lettering on a bottle (User 5428), word choice or images typical to ChatGPT, or a "Sora accent" in how people speak in videos (User 7941). Sometimes, people even identify specific model versions, such as "4o" as shorthand for OpenAI's GPT-4o.

Considering the tf-idf results in Table 2, common words under this sub-code include references to AI or prompting techniques (e.g., *diffusion*, *prompter*), as well as references to specific companies and models (e.g., *dalle*).

Moreover, people sometimes developed guesses about the prompt that could have been used to generate the content. For example, people sometimes identified prompts that they had previously seen produce similar types of content — e.g., observing that the content looks similar to ChatGPT output when it is prompted for "voodoo," "magic," or "occult" (User 4266). Sometimes users experimented directly by attempting to recreate similar output with specific models.

People also mused that if they *could not* imagine a prompt that could have produced the content, that was a signal that content was not AI-generated. For example, User 39411 discussed how they would not know how to prompt an AI model to create a video clip that so accurately reflected real-world physics, thus leading them to a "not AI" conclusion.

External Signals

Beyond content itself, users also leveraged external sources, including understanding where the media originated, when it was created, and how external tools classify it.

Triangulating Information. 5.3% of posted comments triangulate using information from external trustworthy sources (e.g., "lateral reading" (Wineburg and McGrew 2019)), and using either it's presence or absence, as a signal of whether content is likely AI-generated.

Some users attempt to identify and evaluate the original source, such as specific social media accounts or news outlets. Sometimes, the account that produced the content discloses the use of generative AI; other times, other content on the profile dicloses or is interpreted to involve AI use.

Users also find corroborating evidence from third-party sources via search engines (e.g., Google) , news sources (e.g. CNN), or fact-checking services (e.g., BBC Verify). For example, *bbc* appears in the tf-idf results for this strategy (see extended version of the paper (Yeung et al. 2026)). Or, when evaluating a video, one user wrote *"It's real and has been verified by CNN and other news outlets"* (User 45272).

Media Age. 2.1% of posted comments use the age of the content in their decision-making process. Often, this involved determining which AI models were available by that point, and whether during that period such content could be created (see **Model Capabilities**). In some cases, the age of content rules out AI entirely when images were decades old; in other cases, users combine content age with their mental models about model capabilities at that time. For example: "pic is almost 4 years old... anything fine-looking older than 2023 is automatically real" (User 58056). This age was established via a number of methods, including reverse image search, and their existing knowledge of content or events, and the timeframe they occurred within.

Beyond straightforward keywords referring to time (e.g., *dated* or *2013*), Table 2 shows the prevalence of *covid* as a time reference. For instance: *"Not AI, I bought this exact art pre covid for my old place"* (User 107915).

Tool Signals. 1.4% of comments use AI detection tools, watermarks, or ad hoc tools. Specific tool names that appear commonly (Table 2) include *synthid* (watermarking), *tineye* (reverse image search), and several third-party AI detection tools (e.g., *torf.ai*, *aidetective*). We also observed examples of users attempting to use AI chatbots themselves as AI detectors, without any embedded tools.

The relatively limited prevalence of this strategy in our dataset is influenced by the community rules of r/isthisAI: "AI detectors are unreliable and should not be used as evidence for AI or not.". However, r/RealOrAI does not have a similar rule, and r/isthisAI does allow use of watermarks: "SynthID is not an AI detector and therefore allowed."

In some cases, people look for watermarks manually: *"You can see the blurred out sora logo"* (User 11913). In other cases, people look for cryptographic watermarks, such as SynthID: *"If you run this through SynthID it will detect that it's made using nano banana"* (User 16526). Unfortunately, even when people knew about cryptographic watermarks, they sometimes misunderstood them. For instance, one user incorrectly expected a watermark to be detectable by multiple tools: *"I used chat GPT, I used Grok and I even used Deep. Those assistants were not able to tell even after scanning the image. Ford me that's a little inconclusive even though Google confirmed it with the SynthID"* (User 21684).

Lastly, several users reported ad hoc strategies that repurposed existing tools. For example, people often use reverse image search to aid in other strategies: "Reverse image search tells me *"gigantic rat in rice fields' seems to be a bit of an AI video trend"* (User 23545). In one unique example, a commenter evaluated whether a prematurely leaked movie trailer was AI-generated or genuine by uploading the video to YouTube and relying on YouTube's detection of copyrighted content to determine that it was not AI-generated.

Other External. Finally, 0.2% of posted comments relied on sources or knowledge not embedded in the media itself, but not clearly supported by common information infrastructures. For instance, we observed people using content in other comments and other general forum activity as signals in their decision making, such as this individual who noted that the photographer of an animal provided context for the image, saying *"Yeah the photographer who took it mentions it has an injured back leg"* (User 240). The tf-idf results also surfaced *youtube* as a common word in this category.

Perceived Motivations for AIG Content

AI Trends. 1.8% of posted comments consider whether the posted content follows trends and patterns in previously-identified and often viral AI-generated media. For example, AI-generated videos of home surveillance cameras (e.g., Ring doorbells) became popular in 2025 (Marcin 2025), and this made User 7665 believe that all such videos were likely AI: *"Ring/security camera type of video? Check... All of these traits scream AI generated video right away."* Other common trends mentioned were dashcam videos, Halloween videos, and Ghibli-style pictures. For example, one user remarked sarcastically: *"Boy, I'm sure glad all these brand new videos of animals being spooked by Halloween decor next to a full bowl of candy are coming out right BEFORE Halloween"* (User 13547).

Common words describing AI_TRENDS (Table 2) fall into two categories. First, those that express the sheer volume of certain types of content: *popping* (as in, "popping up in my feed"), *flooded*, and *trends*). And second, references to specific trends (e.g., *doorbell* and *porch* in references to AI-generated doorbell camera videos).

Some users tied these trends to specific platforms and target demographics, e.g., Facebook often having AI-generated images of older adults celebrating birthdays alone or of children showing off unrealistically highly-skilled craftwork. For instance: *"Facebook is littered with this slop and everyone's grandparents believes it to be real"* (User 25062).

Motivation For Content. 1.6% of posted comments, people speculate about what incentives someone might post certain AI-generated content, and then use those perceived motivations whether the content likely is, or is not, AI.

Important words here (Table 2) reveal common motivations like perpetrating scams (e.g., *butchering*, as in "pig butchering" scams) and engagement bait (e.g., *ragebait* and *sympathy*). Other motivations discussed include avoiding paying for the rights of commercial images, generating revenue/promote products, and political misinformation or propaganda. For example, in response to a video related to a politically-charged current event, one user posted: *"This one is definitely AI. Be aware that Trump's supporters will pump this out and it will be crucial that we help each other in determining what is real and what is fake"* (User 20404). In another example, one user said, *"I wouldn't trust any images coming from Iranian propaganda, they have been caught using AI images numerous times already"* (User 105113).

Intuition

Lastly, we find that 12.7% of posted comments are decisive, often short statements without additional reasoning. Most of the important words in INTUITION (see tf-idf in the extended version of the paper (Yeung et al. 2026)) suggest confidence: e.g., *fakest*, *painfully*, *christ*, and *1000000*. For instance, one person wrote: *"This is the fakest thing ever. 1000000% AI"* (User 424). People also refer to a different

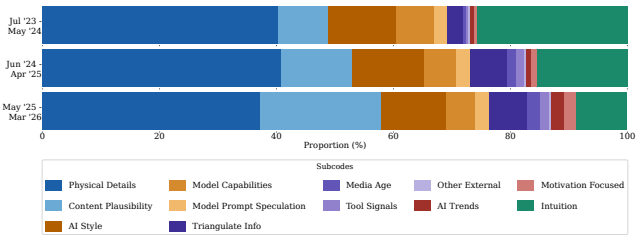

Figure 1: AI Strategies over time: The proportion of each sub-code across our observed time periods.

subreddit, r/isthisacirclejerk, to make the point that something is obviously AI: *"I thought this was the r/isthisacirclejerk sub lol. Extremely ai"* (User 16484).

RQ2: How Strategies Vary

We now investigate: how do people's uses of different AI detection strategies change or vary: over the lifetime of the communities, or across different types of content?

Changes Over Time

In Figure 1, we display the relative frequencies of different strategies used over three 11 month intervals in our dataset: (**p1**) July 2023–May 2024, (**p2**) June 2024–April 2025, and (**p3**) May 2025–March 2026. To evaluate whether the relative proportions of strategies reported in comments vary across time, we conduct a χ^2 test of independence between our time intervals (p1–p2 and p2–p3); and we apply a Bonferroni correction to control for false positives.

Indeed, we find that we find the proportions of reported AI strategies significantly vary between p1 and p2 (χ^2 (11, N=3,165)=65.18; $p_{\text{adj}}<0.025$) as well as p2 and p3 (χ^2 (11, N=320,141)=267.69; $p_{\text{adj}}<0.025$) To evaluate what and how specific strategies change across time, we then conduct a pair-wise series of post-hoc χ^2 tests for each AI strategy within the three compared time periods (e.g. p1–p2, and p2–p3), and again apply a Bonferroni correction among each set of post-hoc tests ($p_{\text{adj}}=0.002$). As shown in Table 3, we find that 7 AI strategies varied significantly in proportions across our evaluated time periods.

Between p1 (July 2023–May 2024) and p2 (June 2024–April 2025), we find that only two strategies significantly varied in prevalence:

1. Intuition decreased, perhaps due to stricter moderator enforcement of community guidelines that prohibit short statements without providing support.
2. Triangulate Info increased, perhaps as users gained experience with AI detection or as the type of content posted changed.

Between p2 (6/2024–4/2025) and p3 (5/2025–3/2026), we find five strategies varied significantly in prevalence:

1. Intuition decreased: As above, this may be due to changing subreddit norms and rules.
2. Physical Details decreased: This relative decline may be in part due to the rise of other strategies (discussed next) that develop as subreddit users gain longitudinal experience with AI-generated content and the surrounding online ecosystems.
3. AI Trends increased: As AI use on social media became more popular and viral successes inspired copycat content, AI trends also appear more often in posts and thus comments. For example, in p3, User 6317 mentions *"Someone is trying to copy the success of the bunnies"* from mid-2025 (Thomas 2025). The extended paper (Yeung et al. 2026) contains additional longitudinal data for trends-related keywords (see "AI Trends Over Time").
4. Content Plausibility increased, perhaps in place of comments expressing a non-justified intuition.
5. Motivation increased: p3 overlaps the 2025 holiday shopping season, during which time there was a growing awareness of AI-generated product scams (Cullinane and Woods 2025), reflected in our data. For example, *"this is totally ai. dont use etsy anymore. there's so much dropshipping and ai bullshit on there meant to just make as much money as possible, disappear, then reappear"* (User 17319).

Variation Across Content

Different strategies may be used for different types of content. Below, we investigate differences between (eventually-concluded) AI versus non-AI content. We provide an additional content-based case study (politics) in the extended version of the paper (Yeung et al. 2026).

Strategies for AI vs. Non-AI Concluded Media. We observed in our qualitative analysis that eleven of the twelve strategies (the exception being AI_STYLE) were used both to justify decisions that something *was* AI-generated as well as that it *was not* AI-generated. However, are there different strategies used more or less often in the two cases?

In November 2025, moderators on r/isthisAI introduced new flairs (specifically, for posts concerning images or videos) that allowed users or moderators to update a post once it has been established that the content in question is or is not AI generated (r/isthisAI Mods 2025). Though these labels exist only for a limited number of posts in our dataset — 843 comments associated with posts that use the "Solved [AI]" flair, and 378 comments associated with posts that use the "Solved [Not AI]" flair — they give us an opportunity to compare strategies that users apply in the two cases.

Indeed, we find significant differences. Via a χ^2 test for independence between posts labeled as "AI" vs "Not AI", we find that the proportion of AI strategies vary significantly (χ^2 (11, N=1,711) = 195.80; $p<0.05$). As shown in Table 4, we evaluate which strategies vary in proportion via a series of pairwise, Bonferroni-corrected, χ^2 tests among every strategy and find that two significantly vary ($p_{\text{adj}}<0.002$).

Two strategies are significantly different for AI compared to non-AI content:

1. AI Style increases (+318.8%): Users typically use AI style to judge content as AI-generated. AI style comments on "Solved [not AI]" posts involve either users making incorrect judgments that something *is* AI-generated, identifying that something is authentic but

	Counts			p1 vs p2			p2 vs p3		
Labels	**p1**	**p2**	**p3**	χ^2	**Delta**	**p-val**	χ^2	**Delta**	**p-val**
PHYSICAL_DETAILS	295	995	118,100	0.060	1.48%	0.807	**14.170**	**-9.10%**	**0.000**
CONTENT_PLAUSIBILITY	62	293	65,654	6.859	42.18%	0.009	**109.218**	**71.60%**	**0.000**
AI_STYLE	86	299	35,654	0.108	4.60%	0.740	2.652	-8.68%	0.103
MODEL_CAPABILITIES	47	132	15,418	0.867	-15.50%	0.352	1.591	-10.55%	0.207
MODEL_PROMPT_SPECULATION	16	62	7,575	0.175	16.58%	0.675	0.213	-6.44%	0.644
TRIANGULATE_INFO	21	152	21,021	**11.785**	**117.77%**	**0.001**	0.474	5.91%	0.491
MEDIA_AGE	3	37	7,151	4.711	271.06%	0.030	5.535	48.01%	0.019
TOOL_SIGNALS	3	33	4,409	3.681	230.95%	0.550	0.002	2.32%	0.964
OTHER_EXTERNAL	2	11	1,372	0.112	65.47%	0.738	0.000	-4.48%	1.000
AI_TRENDS	5	17	6,935	0.000	2.29%	1.000	**24.340**	**212.40%**	**0.000**
MOTIVATION_FOCUSED	4	27	6,787	1.306	103.08%	0.253	**11.725**	**92.50%**	**0.001**
INTUITION	188	375	27,632	**39.884**	**-39.99%**	**0.000**	**135.573**	**-43.57%**	**0.000**

Table 3: Chi-squared post-hoc results, comparing how strategies change across periods (with comparisons between period 1). Rows in bold font indicate statistically significant results, where the *p-value* is less than 0.002 (due to Bonferroni corrections). Delta percentages are based on the proportion of stategy-mentions in a particular period, i.e., the data underlying Figure 1.

Sub-code	**Not AI**	**AI**	χ^2	**Delta**	***p-val***
PHYSICAL_DETAILS	199	437	0.533	-5.2%	1.000
CONTENT_PLAUSIBILITY	95	175	3.569	-20.5%	0.707
AI_STYLE	**20**	**194**	**49.175**	**+318.8%**	**0.000**
MODEL_CAPABILITIES	14	38	0.131	+17.2%	1.000
MODEL_PROMPT_SPECULATION	3	39	9.737	+461.3%	0.022
TRIANGULATE_INFO	**71**	**65**	**32.971**	**-60.5%**	**0.000**
MEDIA_AGE	**47**	**4**	**92.922**	**-96.3%**	**0.000**
TOOL_SIGNALS	10	29	0.198	+25.2%	1.000
OTHER_EXTERNAL	2	2	0.103	-56.8%	1.000
AI_TRENDS	11	46	2.790	+80.6%	1.000
MOTIVATION_FOCUSED	8	26	0.438	+40.3%	1.000
INTUITION	36	140	8.263	+67.9%	0.0485

Table 4: Chi-squared post-hoc comparison of strategies used in "Solved [Not AI]" versus "Solved [AI]" posts (with delta from "Not AI" to "AI" posts). Bolded rows are statistically significant, post-Bonferroni corrections.

modified (*"Verdict: Real but edited with AI"*, User 19924), or identifying non-AI inauthentic content (*"Not AI. Just shitty Photoshop"*, User 70807).

2. Triangulating Info decreases (-60.5%): In both cases, people use external evidence to support their decision; we hypothesize that such evidence may be more common or easier to find for non-AI content.

Discussion and Conclusion

Finally, we step back and reflect on (1) lessons about the mental models that we see people developing around AI-generated content detection, and (2) what our findings suggest about potential approaches to support uses in this task.

Mental Models for Detecting AI Content

Stepping back from individual AI detection strategies, we draw several higher-level lessons about the mental models that commenters in our dataset have developed.

Content-focused strategies dominate strategies that rely on specific external signals. In our dataset, the most commonly used strategies consider only the specific piece of content under evaluation, often combined with a user's evolving mental model of what AI-generated content "looks like" or what models are capable for producing. By contrast, strategies that rely on external trustworthy sources or tools are much more rare. This is not necessarily a good or a bad thing (indeed, AI detection tools are not robust, as we have discussed, and appear less in the data because of one of the subreddit's rules). Still, it suggests a potential opportunity for an increased role of authoritative external tools and/or resources; though it also suggests potential challenges for such external resources, since users may prefer the lower-effort, content-focused heuristics.

Many popular strategies are detail-oriented or model-specific, and may not be robust to AI model evolution. Our findings suggest that people are developing mental models of what AI-generated content looks like, built on personal experiences and beliefs developed over time. People often focus on details to identify AI-generated content, or to justify their intuition — focusing on physical details (e.g., hands), the plausibility of certain aspects of the content (e.g., a child's behavior), or specific indications of AI style (e.g., water texture). These details are then intersected with the assumptions that people have developed about the capabilities of AI models. While these strategies can be very effec-

tive in some cases, they are likely not robust to evolutions in AI models; and indeed, in some ways this is a fundamental "arms race". Thus, such detail-oriented strategies should be complemented with others, which rely on external sources and/or a broader understanding of the ecosystem.

Some heuristic strategies rely on broader context or experience that particularly vulnerable users may not have. Our results suggest increasing use of strategies that rely not just on the detailed characteristics of a piece of content, but on a broader understanding of the online ecosystem: what motivates people to post AI-generated content, and what types of AI-generated memes are currently going viral. While these may be effective strategies for users who are very "online", such as Reddit commenters and some of this paper's authors, they rely on exposure and experience over time that many people may not have (e.g., older adults, or less tech-savvy users). The consequence is that these people may be more vulnerable to AI-generated content, which can be problematic if that content also spreads mis/disinformation (e.g., political or health). Moreover, because these types of strategies rely on accumulated experience across many pieces of content, they may be hard to teach.

Mental models are socially constructed, and there is a risk of false confidence from unreliable tools or strategies. Particularly in our dataset, where users interact specifically for the purpose of discussing AI-generated content detection, the emerging strategies and underlying mental models are socially constructed and reinforced. The same is true in other contexts where users engage in collective sensemaking, e.g., in the comments on a Facebook post. This can be helpful to people, as they learn and spread information about how to recognize AI-generated content. However, this can also have unintended negative effects, such as when the strategies people share are outdated relative to model capabilities, when AI detection tools are used and recommended but not actually effective, or when incorrect strategies may systematically categorize 'non-AI' media as 'AI' according to held biases of others' identities (Mink et al. 2024). The resulting reliance on such strategies or tools can, in turn, risk creating false confidence that a particular piece or type of content is authentic, or that authentic content is fictitious.

Lessons for Existing Solutions

We reflect on what our findings suggest about proposed mechanisms for helping users identify AI-generated content.

Watermarking. Prior academic work has discussed alternative methods for AI detection, including both watermarking and other cryptographic solutions. These solutions provide several benefits, including clear provenance information and the opportunity to rely on detection tools rather than individual human knowledge. However, in addition to other limitations with such approaches (discussed in the Related Work), our work highlights important usability challenges. For instance, we observed people who knew that SynthID (Google's watermarking) indicated generative AI use, but **misunderstood how to check for it**, e.g., thinking that they could ask another service (such as ChatGPT) to check for the Google watermark. Thus, we find that it is not enough to develop the technical capabilities the watermarking and cryptographic techniques enable. It is also important to consider the **usability** of these tools as well, both in the interface that companies present to people, as well as the information they release alongside these tools.

AI Disclosures. Existing research has demonstrated limitations of AI content disclosures that are starting to be deployed on some platforms, e.g., that users may not fully understand what the disclosures mean (Holtervennhoff et al. 2026; Gamage et al. 2025; Møller et al. 2026). Our findings show that people *are* sometimes leveraging these disclosures, e.g., looking for cues like AI disclosures in an account profile or in an account's other content to assess a piece of content under scrutiny. However, this process is ad hoc and disclosures currently vary widely across platforms. This lack of standardization may muddy waters for users. Rather than reinvent the wheel every time when considering disclosures, **we urge companies to collaborate and unify disclosures** in a way that prioritizes user understanding — e.g., via the existing C2PA coalition (C2PA 2026).

Consumer Education. Though the burden should certainly not fall on users alone to identify AI-generated content, consumer education and guidance can play a role in helping people develop accurate mental models and effective strategies. Our work highlights the importance of doing **human-centered research when developing and releasing guidance** to understand the strategies people are already using, which strategies have staying power, which strategies are misunderstood, and how those strategies are succeeding or failing. For example, we find that people often consider the details of specific content and rely on their assumptions of AI model capabilities — though these strategies can be effective, education for users should also convey that detail-focused strategies can become outdated as models improve. Echoing significant past work in the broader space of online mis/disinformation, we encourage the development of guidance and education that emphasizes strategies that will work even as AI models evolve, and do not disproportionately harm groups of people along the way (Mink et al. 2024) e.g., lateral reading, triangulation, and relying on trustworthy external sources (Caulfield 2019; Kampen 2025).

Future Work

Leveraging our labeled dataset[2], future work might: investigate how strategies vary across media type or content type; dive deeply into political content in particular (labeled with a "flair" in r/isthisAI); and investigate the effectiveness of strategies (e.g., by leveraging the "[GUESS]" posts on r/RealOrAI that provide a ground-truth answer).

More broadly, future work might complement our investigation with studies of other user populations or datasets; with in-depth interviews to dig deeper into people's AI detection mental models; with analyses of publicly-available guides for identifying AI-generated content; and with investigations of the impacts of AI detection tools and platform labels on people's skills and perceptions.

[2]https://agent-security.cs.washington.edu/is-this-ai.html

Acknowledgments

We are very grateful to the following people for their feedback and conversations in the development of this work and paper: Natalie Grace Brigham, Maddie Burbage, Ian Chang, Elise Dorough, Joe Eckert, Yael Eiger, Kshitish Ghate, Blaine Hoak, Rachel Hong, David Kohlbrenner, Rachel McAmis, Chris Nagle, Basia Radka, Les Sessoms, Michael Tompkins, Henry Wong. Thank you especially to Natalie Grace Brigham for help with qualitative coding. This work was supported in part by the U.S. National Science Foundation under Award #2114230 and by a Microsoft Grant for Customer Experience Innovation.

References

Anthropic. 2026. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-05-17.

Azungah, T. 2018. Qualitative Research: Deductive and Inductive Approaches to Data Analysis. *Qualitative Research Journal*, 18(4): 383–400.

Bouma-Sims, E.; Hassan, H.; Nisenoff, A.; Cranor, L. F.; and Christin, N. 2024. "It was honestly just gambling": Investigating the Experiences of Teenage Cryptocurrency Users on Reddit. In *Symp. on Usable Privacy and Security*, 333–352.

Bouma-Sims, E.; Lanyon, M.; and Cranor, L. F. 2025. "Is this a scam?": The Nature and Quality of Reddit Discussion about Scams. In *Proc. of the 2025 ACM SIGSAC Conf. on Computer and Communications Security*, 2444–2458.

Bray, S. D.; Johnson, S. D.; and Kleinberg, B. 2023. Testing human ability to detect 'deepfake'images of human faces. *Journal of Cybersecurity*, 9(1): tyad011.

C2PA. 2026. Coalition for Content Provenance and Authenticity. https://c2pa.org/.

Cai, S.; and Cui, W. 2023. Evade ChatGPT Detectors via A Single Space. arXiv:2307.02599.

Caulfield, M. 2019. SIFT (The Four Moves). https://hapgood.us/2019/06/19/sift-the-four-moves/. Accessed: 2026-05-17.

Christ, M.; Gunn, S.; and Zamir, O. 2024. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, 1125–1139. PMLR.

Chu, H.; and Liu, S. 2024. Can AI tell good stories? Narrative transportation and persuasion with ChatGPT. *Journal of Communication*, 74(5): 347–358.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Cooke, D.; Edwards, A.; Barkoff, S.; and Kelly, K. 2025. As Good as a Coin Toss: Human Detection of AI-Generated Content. *Communications of the ACM*, 68(10): 100–109.

Cullinane, A.; and Woods, R. 2025. Meta accused of letting AI sellers 'run rampant'. https://www.bbc.com/news/articles/c4gpz90d4q0o.

Dathathri, S.; See, A.; Ghaisas, S.; Huang, P.-S.; McAdam, R.; Welbl, J.; Bachani, V.; Kaskasoli, A.; Stanforth, R.; Matejovicova, T.; et al. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035): 818–823.

Diel, A.; Lalgi, T.; Schröter, I. C.; MacDorman, K. F.; Teufel, M.; and Bäuerle, A. 2024. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16: 100538.

Elkhatat, A. M.; Elsaid, K.; and Almeer, S. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1): 17.

Ernst, M.; Rupp, F.; and Simbeck, K. 2025. You Can't Detect Me! Using Prompt Engineering to Generate Undetectable Student Answers. In *CSEDU (2)*, 304–311.

Farid, H. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4).

Fiesler, C.; and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1).

Fu, X.; Liu, S.; Xu, Y.; Lu, P.; Hu, G.; Yang, T.; Anantasagar, T.; Shen, C.; Mao, Y.; Liu, Y.; Shah, K.; Lee, C. U.; Choi, Y.; Zou, J.; Roth, D.; and Callison-Burch, C. 2025. Learning Human-Perceived Fakeness in AI-Generated Videos via Multimodal LLMs. arXiv:2509.22646.

Gamage, D.; Sewwandi, D.; Zhang, M.; and Bandara, A. K. 2025. Labeling synthetic content: User perceptions of label designs for AI-generated content on social media. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, 1–29.

Gibson, C.; Olszewski, D.; Brigham, N. G.; Crowder, A.; Butler, K. R.; Traynor, P.; Redmiles, E. M.; and Kohno, T. 2025. Analyzing the {AI} nudification application ecosystem. In *34th USENIX Security Symposium (USENIX Security 25)*, 1–20.

Gotoman, J.; Luna, H.; Sangria, J.; Santiago Jr, C.; and Barbuco, D. 2025. Accuracy and reliability of AI-generated text detection tools: a literature review. *American Journal of IR*, 4(0): 1–9.

Heitmann, A. 2024. arctic_shift. https://github.com/ArthurHeitmann/arctic_shift. GitHub repository, accessed 2026-05-17.

Hoffman, B. 2024. First Came "Spam." Now, With A.I., We've Got "Slop". https://www.nytimes.com/2024/06/11/style/ai-search-slop.html.

Höltervennhoff, S.; Ricker, J.; Raphael, M. M.; Schwedes, C.; Weil, R.; Fischer, A.; Holz, T.; Schönherr, L.; and Fahl, S. 2026. "That's another doom I haven't thought about": A User Study on AI Labels as a Safeguard Against Image-Based Misinformation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–32.

Holtervennhoff, S.; Ricker, J.; Raphael, M. M.; Schwedes, C.; Weil, R.; Fischer, A.; Holz, T.; Schönherr, L.; and Fahl, S. 2026. "That's another doom I haven't thought about":

A User Study on AI Labels as a Safeguard Against Image-Based Misinformation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.

Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; and McAuley, J. 2021. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3348–3357.

JeremyFindsAI. 2026. Instagram profile: JeremyFindsAI. https://www.instagram.com/jeremyfindsai/. Accessed: 2026-05-07.

Kampen, K. V. 2025. Library Guides: Evaluating Resources and Misinformation: The SMART Check. https://guides.lib.uchicago.edu/c.php?g=1241077&p=9082345.

Kennedy, B.; Yam, E.; Kikuchi, E.; Pula, I.; and Fuentes, J. 2025. How Americans View AI and Its Impact on People and Society. Accessed: 2026-05-02.

Kraemer, H. C.; Kupfer, D. J.; Clarke, D. E.; Narrow, W. E.; and Regier, D. A. 2012. DSM-5: How Reliable Is Reliable Enough? *American Journal of Psychiatry*, 169(1): 13–15.

Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; and Iyyer, M. 2023. Paraphrasing Evades Detectors of AI-generated Text, but Retrieval is an Effective Defense. arXiv:2303.13408.

Layton, S.; Medeiros, B. B. P.; Butler, K.; and Traynor, P. 2026. AI Wrote My Paper and All I Got Was This False Negative: Measuring the Efficacy of Commercial AI Text Detectors. In *IEEE Symposium on Security & Privacy*.

Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; and Zou, J. 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7): 100779.

Lima, G.; Grgić-Hlača, N.; and Redmiles, E. M. 2025. Public Opinions About Copyright for AI-Generated Art: The Role of Egocentricity, Competition, and Experience. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–32. ACM.

Liu, A.; Pan, L.; Lu, Y.; Li, J.; Hu, X.; Zhang, X.; Wen, L.; King, I.; Xiong, H.; and Yu, P. 2024. A Survey of Text Watermarking in the Era of Large Language Models. *ACM Comput. Surv.*, 57(2).

Lu, N.; Liu, S.; He, R.; Wang, Q.; Ong, Y.-S.; and Tang, K. 2024. Large Language Models can be Guided to Evade AI-Generated Text Detection. arXiv:2305.10847.

Mai, K. T.; Bray, S.; Davies, T.; and Griffin, L. D. 2023. Warning: Humans cannot reliably detect speech deepfakes. *Plos one*, 18(8): e0285333.

Marcin, T. 2025. AI-generated animals in fake surveillance videos are fooling the internet. https://mashable.com/article/ai-generated-animals-footage-fake.

Mink, J.; Luo, L.; Barbosa, N. M.; Figueira, O.; Wang, Y.; and Wang, G. 2022. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *31st USENIX Security Symposium (USENIX Security 22)*, 1669–1686. Boston, MA: USENIX Association. ISBN 978-1-939133-31-1.

Mink, J.; Wei, M.; Munyendo, C. W.; Hugenberg, K.; Kohno, T.; Redmiles, E. M.; and Wang, G. 2024. It's Trying Too Hard To Look Real: Deepfake Moderation Mistakes and Identity-Based Bias. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Møller, A. G.; Romero, D. M.; Jurgens, D.; and Aiello, L. M. 2026. The impact of generative AI on social media: An experimental study. *Scientific Reports*.

Müller, N. M.; Pizzi, K.; and Williams, J. 2022. Human perception of audio deepfakes. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia*, 85–91.

Naitali, A.; Ridouani, M.; Salahdine, F.; and Kaabouch, N. 2023. Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Computers*, 12(10): 216.

Nath, S.; Huang, C.-Y.; Ganapathi, A.; Thimmaraju, K.; Mink, J.; and Ahn, G.-J. 2026. Like a Hammer, It Can Build, It Can Break: Large Language Model Uses, Perceptions, and Adoption in Cybersecurity Operations on Reddit. arXiv:2604.09998.

Oak, R.; and Shafiq, Z. 2025. Victims, Vigilantes, and Advice Givers: An Analysis of {Scam-Related} Discourse on Reddit. In *SOUPS 2025*, 57–71.

Peckham, J. 2025. Sora 2 Now Lets You Make 15-Second Clips, 25 Seconds for $200 Pro Users. Accessed: 2026-05-20.

Reddit. 2026a. r/AIorFake. https://www.reddit.com/r/AIorFake/. Accessed: 2026-05-17.

Reddit. 2026b. r/AI_or_Real. https://www.reddit.com/r/AI_or_Real. Accessed: 2026-05-17.

r/isthisAI Mods. 2025. Flair has been added along with a SOLVED flair. https://www.reddit.com/r/isthisAI/comments/1ox0uk2/flair_has_been_added_along_with_a_solved_flair/.

Roca, T.; Roman, A. C.; Vega, J. T.; Duarte, M.; Wang, P.; White, K.; Misra, A.; and Ferres, J. L. 2025. How good are humans at detecting AI-generated images? Learnings from an experiment. arXiv:2507.18640.

Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2025. Can AI-Generated Text be Reliably Detected? arXiv:2303.11156.

Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5): 513–523.

Shaib, C.; Chakrabarty, T.; Garcia-Olano, D.; and Wallace, B. C. 2025. Measuring AI" Slop" in Text. *arXiv preprint arXiv:2509.19163*.

Thomas, R. 2025. AI Bunnies on Trampoline Causing Crisis of Confidence on TikTok. https://www.404media.co/ai-bunnies-on-trampoline-causing-crisis-of-confidence-on-tiktok/.

Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šigut, P.; and Waddington, L. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1).

Wei, M.; Consolvo, S.; Kelley, P. G.; Kohno, T.; Matthews, T.; Meiklejohn, S.; Roesner, F.; Shelby, R.; Thomas, K.; and Umbach, R. 2024. Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse. In *33rd USENIX Security Symposium (USENIX Security 24)*, 4391–4408. Philadelphia, PA: USENIX Association. ISBN 978-1-939133-44-1.

Wineburg, S.; and McGrew, S. 2019. Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record*, 121(11): 1–40.

Yeung, C.; Weld, G.; Mink, J.; and Roesner, F. 2026. Is This AI? Longitudinal Analysis of Strategies Used for AI Detection on Two Subreddits. arXiv:2606.22689.

Zhao, X.; Gunn, S.; Christ, M.; Fairoze, J.; Fabrega, A.; Carlini, N.; Garg, S.; Hong, S.; Nasr, M.; Tramer, F.; Jha, S.; Li, L.; Wang, Y.-X.; and Song, D. 2025. SoK: Watermarking for AI-Generated Content. arXiv:2411.18479.

Zhao, X.; Zhang, K.; Su, Z.; Vasan, S.; Grishchenko, I.; Kruegel, C.; Vigna, G.; Wang, Y.-X.; and Li, L. 2024. Invisible Image Watermarks Are Provably Removable Using Generative AI. arXiv:2306.01953.

Appendices

Ethics and Adverse Impact Statement

We describe here several ethical considerations that arose in the design and execution of our research:

1. Our University's IRB does not consider studies like this one to be federally-regulated human subjects research, because they work with public data that does not (typically) contain personally-identifiable or other sensitive information. Nevertheless, we acknowledge that although we draw our dataset from a public source (publicly-available Reddit comments), there are still potential ethical concerns with using public data for research in ways that is unexpected by or potentially harmful to the people whose data is being used (Fiesler and Proferes 2018). Though the population of users that we study here, and the topics of the subreddits we study, are not particularly vulnerable or sensitive, we nevertheless take several precautions. In the paper, we do not name specific users or link to specific posts or comments, instead using anonymized user identifiers. When we make our labeled dataset publicly available, we will likewise exclude reddit username information. Finally, we did our best to remove posts and comments from our dataset (collected with ArcticShift) that were since deleted from the live Reddit site. We also verified that neither r/RealOrAI nor r/isthisAI have community guidelines in place that prohibit academic research from being conducted.
2. There are also ethical considerations with using LLMs to label data, since this may (for example) result in the unexpected downstream use of sensitive or participant data for model training. We selected the option to disallow Claude from using our data for training, and we used Anthropic's API to batch-send data, which (according to the company's documentation) is not used for training or improving models. Moreover, it is likely that public Reddit data has already been ingested by many companies for model training.
3. In terms of adverse impacts, we recognize that AI companies or AI-generated content creators who are acting in bad faith (or simply want to improve the "quality" of their content by some metrics) could use our findings to make AI-generated content harder to detect. However, this is an ongoing arms race that is likely to be minimally impacted by our paper. Moreover, the data upon which we base our analyses are all already public on Reddit. Ultimately, we hope that this work will have positive impacts that outweigh its contributions to this arms race, in laying a foundation for future work that improves tools and education for consumers.

Author Positionality Statement

All of the co-authors of this paper are based in the United States (one immigrated as a child from Western Europe), where our outlook is shaped by Western cultures and Western media systems. Of the four co-authors, none of them are active community members (i.e., moderate or post or comment) on either r/RealOrAI or r/isthisAI, nor on any similar communities. The authors' combined areas of research experience cover computer security and privacy (including usable security), internet measurement, online communities, mis/disinformation, and perceptions of synthetic media.