The Impact of AI-Generated Text on the Internet

Jonas Dolezal[1], **Sawood Alam**[2], **Mark Graham**[2] **and Maty Bohacek**[3]
[1]Imperial College London, [2]Internet Archive, [3]Stanford University

The proliferation of AI-generated and AI-assisted text on the internet is feared to contribute to a degradation in semantic and stylistic diversity, factual accuracy, and other negative developments (sometimes subsumed under the "Dead Internet Theory"). What has hindered answering these questions is that it has not been understood just how much of the internet is actually AI-generated or AI-edited. To this end, we construct a representative sample of websites published on the internet between 2022 and 2025 using the Internet Archive, and apply a state-of-the-art AI text detector on them. We find that by mid-2025, roughly 35% of newly published websites were classified as AI-generated or AI-assisted, up from zero before ChatGPT's launch in late 2022. We also find statistically significant evidence for some of the identified hypotheses; for example, that increases in AI-generated text on the internet correlate negatively with semantic diversity and positively with the prevalence of positive sentiment. We do not, however, find statistically significant evidence supporting the hypothesis that an increased rate of AI-generated text on the internet decreases factual accuracy or stylistic diversity. Notably, this diverges from public perception, which we measure in a user study, where the majority of US adults turned out to believe in all four of the above-mentioned hypotheses. Individuals who do not use AI or use it infrequently tend to believe in these negative impacts more than those who use it frequently; similarly, individuals who hold negative views of AI tend to believe in these hypotheses more than those with favorable views of the technology.

Keywords: Artificial Intelligence, Large Language Models, AI-Generated Text, Internet, Online Discourse

Ever since ChatGPT first made large language models (LLMs) available to the wider public in 2022, which was followed by mass adoption, there have been concerns about the impact of AI-generated text (as well as AI-generated content in other modalities) on the internet and online discourse (Ferrara, 2026; Muzumdar et al., 2025). Specifically, many known limitations and failure modes of LLMs, including factual hallucinations (Huang et al., 2025), sycophancy (Malmqvist, 2025), verbosity (Saito et al., 2023), and more, have raised concerns that unchecked proliferation of such content could reduce the overall quality of internet content (Shumailov et al., 2024; Xing et al., 2025). These hypotheses are sometimes subsumed under the "Dead Internet Theory," which they loosely expand, but which, on its own, predates the widespread use of LLMs (Muzumdar et al., 2025). These hypotheses have been difficult to verify, primarily because there is limited understanding of how much internet content is actually AI-generated (Santy et al., 2025; Spennemann, 2025). In this paper, we attempt to address these questions. We concern ourselves only with LLM-generated text, leaving other modalities for future work, and use LLM-generated and AI-generated interchangeably.

Verifying the above-mentioned hypotheses about AI-generated text on the internet is difficult for several reasons. First, obtaining representative samples of text from the internet is challenging (Thompson, 2024). As a result, existing analyses have mainly focused on restricted subsets of the internet, such as specific social media platforms (La Cava et al., 2025; Matatov et al., 2024; Paredes et al., 2025; Sun et al., 2025) or news sources (Russell et al., 2025), scientific publishing (Kobak et al., 2025; Liang et al., 2024), software repositories (Daniotti et al., 2026), or translations (Thompson et al., 2024). Second,

Corresponding author(s): maty@stanford.edu ; *Code & Data:* https://ai-on-the-internet.github.io

detecting AI-generated content is itself a difficult problem with many known challenges (Dawkins et al., 2025; Fraser et al., 2025; Sadasivan et al., 2025). To date, AI detection for image and video content has been believed to be more accurate than for text (Wu et al., 2025). Thus, the few existing inquiries in this domain have predominantly focused on image and video modalities rather than text (Matatov et al., 2024). To the best of our knowledge, no prior study has analyzed the impact of AI-generated text on the internet as a whole. We attempt to do so here.

We pose the following research questions: (RQ1) What are the generally held beliefs about the impacts of AI-generated text on the internet among adults in the United States? (RQ2) How many websites on the internet, published between 2022 and 2025, contain AI-generated text? (RQ3) Are the hypotheses about the impact of AI-generated text on the internet from RQ1 correct? To answer RQ1, we conduct a study of a representative sample of US adults to test beliefs about six hypotheses identified by the research team through exploratory environmental scanning and thematic analysis of online discourse. We then sample representative sets of websites from the Internet Archive published between 2022 and 2025 and detect AI-generated text using the detector that performed best in our independent evaluation. Finally, we conduct a series of quantitative experiments to evaluate the hypothesized impacts against this data.

1. Results

We found that the prevalence of AI-generated and AI-assisted websites has been growing since the launch of ChatGPT in November 2022. By the first half of 2025, as much as 35% of websites uploaded to the internet in a given month were AI-generated or AI-assisted. The share of AI-generated and AI-assisted websites over time is shown in Figure 1.

We next review each of the six hypotheses about the impact of AI-generated and AI-assisted text on the internet, presenting the results of our participant survey as well as a quantitative analysis against internet data from the Internet Archive. The aggregate AI likelihood score represents the likelihood of text to be AI-generated and AI-assisted in a given sample.

The Semantic Contraction Hypothesis (Hyp. 1). The statement of this hypothesis is the following: "As AI text becomes more common on the internet, the range of unique ideas and diverse viewpoints shrinks." The results are shown in Figure 2. 60.9% of respondents lean towards agreement with this statement, 11.7% are neutral, and 27.4% lean towards disagreement. This has been translated into the following measurable signal: the average pairwise cosine similarity of semantic embeddings within a monthly sample. If the hypothesis is correct, we posit this signal would be positively correlated with the aggregate AI likelihood score. The null hypothesis (H_0, i.e., the correlation $\rho = 0$) was rejected ($\rho = 0.47$, $p = 0.004$), confirming the hypothesis. Across all evaluated months, the average semantic similarity between websites predicted to be AI-generated or AI-assisted was 33% higher than that of non-AI websites (semantic similarity scores of 0.0701 vs. 0.0526, respectively).

The Truth Decay Hypothesis (Hyp. 2). The statement of this hypothesis is the following: "As AI content becomes more common on the internet, I am encountering factually incorrect information and hallucinations more frequently." 75.1% of respondents lean towards agreement with this statement, 14.3% are neutral and 20.6% lean towards disagreement. This has been translated into the following measurable signal: the average rate of factually incorrect statements within a monthly sample. If the hypothesis is correct, we posit this signal would be positively correlated with the aggregate AI likelihood score. The null hypothesis (H_0, i.e., the correlation $\rho = 0$) was not rejected ($\rho = -0.19$, $p = 0.27$).

The Positivity Shift Hypothesis (Hyp. 3). The statement of this hypothesis is the following: "As AI content becomes more common on the internet, online writing feels increasingly sanitized and artificially cheerful." The results are shown in Figure 3. 72.0% of respondents lean towards

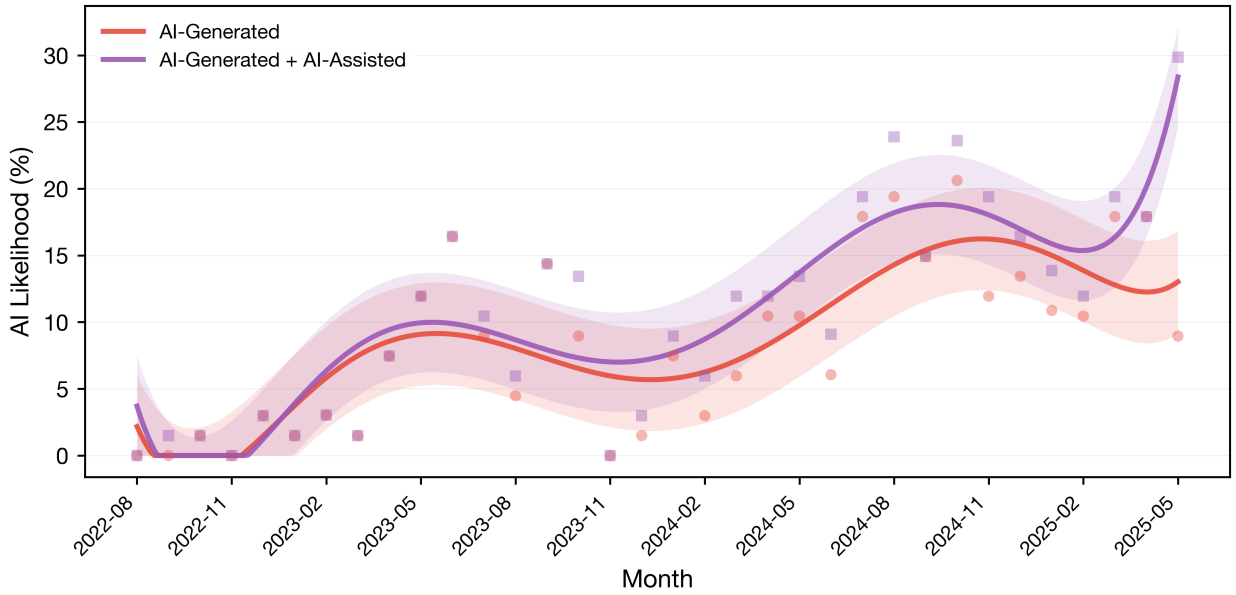


Figure 1 | **AI-generated Text on the Internet from Mid-2022 to Mid-2025.** The figure shows the proportion of websites classified as fully AI-generated (red) and AI-generated or AI-assisted (purple) based on Pangram v3 detection applied to representative samples obtained from the Internet Archive. Curves represent smoothed estimates.

agreement with this statement, 15.3% are neutral and 12.7% lean towards disagreement. This has been translated into the following measurable signal: the rate of positive documents (i.e., classified as positive by sentiment analysis) within a monthly sample. If the hypothesis is correct, we posit this signal would be positively correlated with the aggregate AI likelihood score. The null hypothesis (H_0, i.e., the correlation $\rho = 0$) was rejected ($\rho = 0.56$, $p = 0.0003$), confirming the hypothesis. Across all evaluated months, the average positive sentiment score of AI-generated or AI-assisted was 107% higher than that of non-AI websites (sentiment scores of 0.7042 vs. 0.3400, respectively).

The Epistemic Island Hypothesis (Hyp. 4). The statement of this hypothesis is the following: "As AI content becomes more common on the internet, articles are increasingly providing answers without including links to external sources." 69.9% of respondents lean towards agreement with this statement, 18.7% are neutral and 11.4% lean towards disagreement. This has been translated into the following measurable signal: the density of outbound link tags. If the hypothesis is correct, we posit this signal would be inversely correlated with the aggregate AI likelihood score. The null hypothesis (H_0, i.e., the correlation $\rho = 0$) was not rejected ($\rho = -0.12$, $p = 0.48$).

The Entropy Dilution Hypothesis (Hyp. 5). The statement of this hypothesis is the following: "As AI content becomes more common on the internet, content is becoming significantly longer in word count while having lower semantic density." 60.7% of respondents lean towards agreement with this statement, 13.8% are neutral and 25.5% lean towards disagreement. This has been translated into the following measurable signal: the Gzip compression ratio calculated as the fraction of raw document size over compressed size. If the hypothesis is correct, we posit this signal would be positively correlated with the aggregate AI likelihood score. The null hypothesis (H_0, i.e., the correlation $\rho = 0$) was not rejected ($\rho = -0.02$, $p = 0.89$).

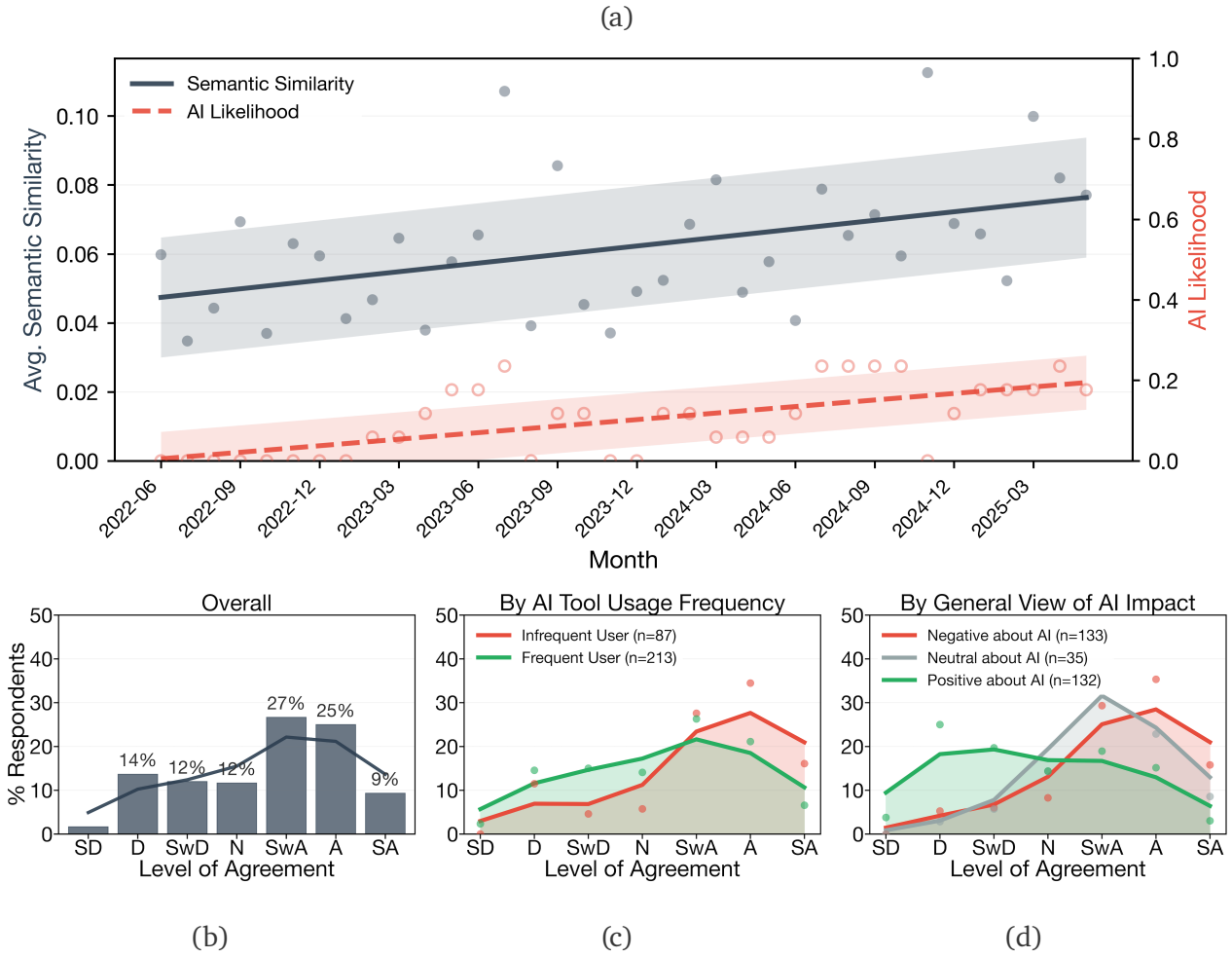


Figure 2 | **Results for Hyp. 1: Semantic Contraction.** The figure shows results for the Semantic Contraction Hypothesis from the participant study (RQ1) and quantitative analysis of randomly sampled websites from the Internet Archive (RQ3). In (a), the average pairwise cosine similarity of semantic embeddings is plotted against AI Likelihood score, which combines the rate of AI-generated and AI-assisted samples, as detected by Pangram v3 ($\rho = 0.47$, $p = 0.004$). The overall results of the participant study are shown in (b), with responses ranging from Strongly Disagree (SD) to Strongly Agree (SA). These are broken down by AI usage frequency in (c) and general view of AI impact in (d).

The Stylistic Monoculture Hypothesis (Hyp. 6). The statement of this hypothesis is the following: "As AI content becomes more common on the internet, distinct individual writing styles are disappearing in favor of a generic, uniform voice." 83.0% of respondents lean towards agreement with this statement, 9.4% are neutral and 7.6% lean towards disagreement. This has been translated into the following measurable signal: the average pairwise cosine similarity of document writing style embeddings within a monthly sample. If the hypothesis is correct, we posit this signal would be positively correlated with the aggregate AI likelihood score. The null hypothesis (H_0, i.e., the correlation $\rho = 0$) was not rejected ($\rho = 0.24$, $p = 0.17$).

Impact of AI Usage and Favorability. For all tested hypotheses, participants who use AI infrequently were more likely to believe that the negative impact of AI-generated content on the internet is real than those who use AI tools regularly, with a pooled agreement rate of 88.3% versus 76.2% among non-neutral respondents (a difference of 12.1 percentage points). Similarly, participants with a less favorable view of AI's gen-

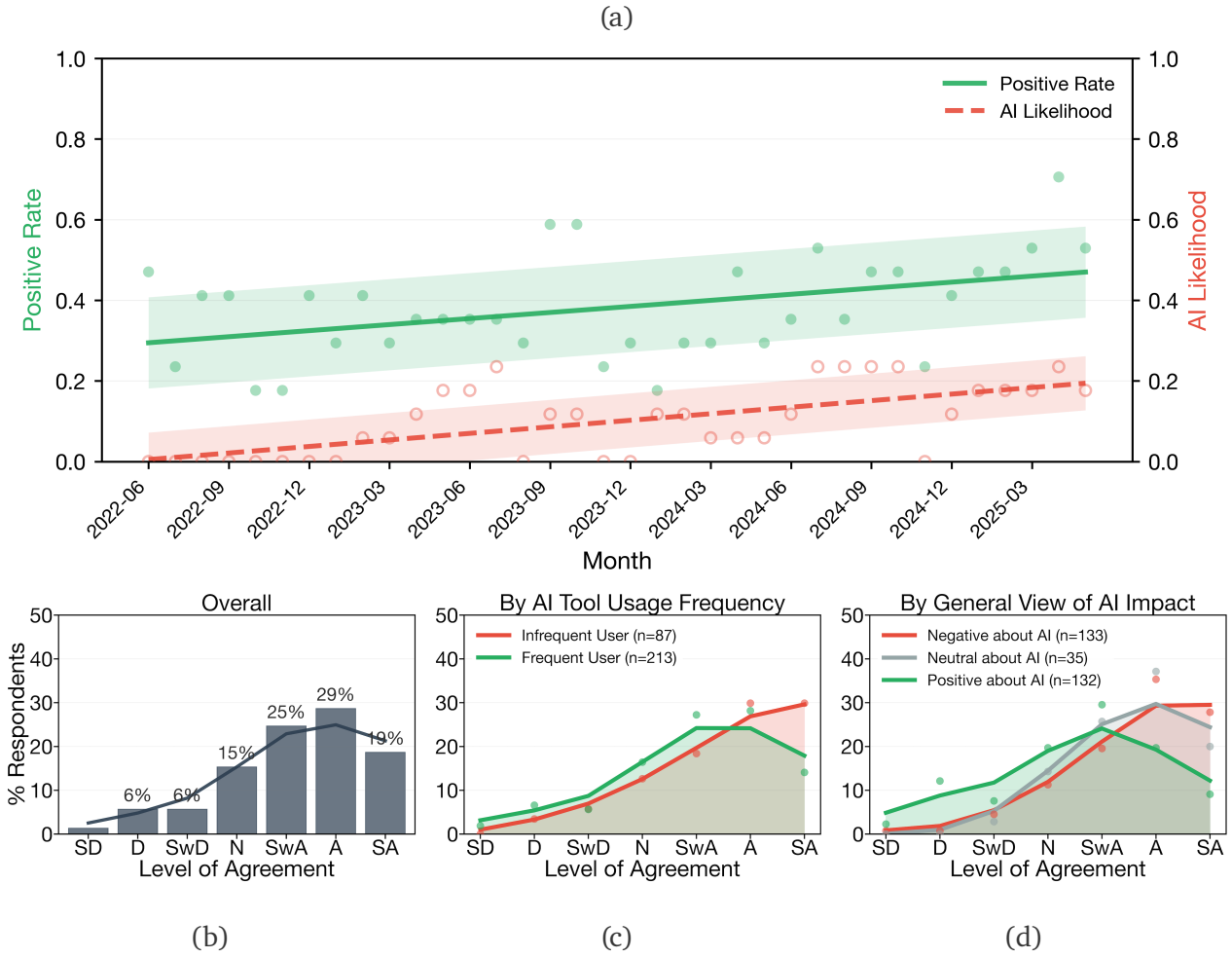


Figure 3 | **Results for Hyp. 3: Positivity Shift.** The figure shows results for the Positivity Shift Hypothesis from the participant study (RQ1) and quantitative analysis of randomly sampled websites from the Internet Archive (RQ3). In (a), the rate of positive documents (classified by sentiment analysis) is plotted against AI Likelihood score, which combines the rate of AI-generated and AI-assisted samples, as detected by Pangram v3 ($\rho = 0.56$, $p = 0.0003$). The overall results of the participant study are shown in (b), with responses ranging from Strongly Disagree (SD) to Strongly Agree (SA). These are broken down by AI usage frequency in (c) and general view of AI impact in (d).

eral impact on society were more likely to believe in the negative impact of AI-generated content on the internet than those with a favorable or neutral view of the technology's societal impact, with a pooled agreement rate of 91.3% versus 71.1% (a difference of 20.2 percentage points).

2. Discussion

Our study shows a shift in the composition of the open web, estimating that as much as 35% of newly published websites by mid-2025 have been AI-generated or AI-assisted. Notably, we find a divergence between the impacts of this shift on online discourse and the public perception of this phenomenon. While our survey (RQ1) reveals a public concern about systemic truth decay (Hyp. 2) and stylistic homogenization (Hyp. 6) as a result of AI-generated text proliferation, our web-scale analysis (RQ3) does not yield statistically significant evidence of macro-level degradation in factual accuracy or a strict stylistic monoculture.

This divergence suggests that the immediate threat to online discourse may be of an epistemic nature rather than purely factual. As AI-generated text becomes ubiquitous and indistin-

guishable from human writing (Chein et al., 2024; Jakesch et al., 2023; Porter and Machery, 2024), users may discount the credibility of all online information (operating on the principle of "reality apathy" or, when used maliciously, the "liar's dividend") (Altay and Gilardi, 2024; Chesney and Citron, 2019; Liu et al., 2025; Schiff et al., 2025). This could potentially alter online news consumption patterns, driving users toward more insular information ecosystems (Jacob et al., 2025; Kitchens et al., 2020). If true, infrequent users of AI tools and general skeptics of the technology would likely be most affected, since we find they harbor a deeper concern about the negative impact of AI-generated text on the internet than their counterparts.

Rather than an explosion of falsehoods, the footprint of AI proliferation on the internet manifests primarily as semantic contraction (Hyp. 1) and an artificial positivity shift (Hyp. 3). The increase in semantic similarity among AI-generated text compared to human-written text implies that the online Overton window may be narrowing, as LLMs optimize for outputs that fall within a more constrained distribution closer to the average of the training distribution (Agarwal et al., 2025; Dohmatob et al., 2024a; Zhang et al., 2025). Simultaneously, the increase in positive sentiment, symptomatic of the sycophantic and overoptimistic nature of existing LLMs (Chen et al., 2025; Malmqvist, 2025; Sharma et al., 2023), implies that the discourse may be becoming more sanitized. Pluralistic online engagement relies on friction, debate, and the processing of diverse societal realities, including the negative ones (Lasser and Poechhacker, 2025; Mouffe, 1999); an environment flooded with cheerful, homogenized text may marginalize human dissent (Daryani et al., 2026; Oh and Downey, 2025). This semantic contraction could quietly pacify or homogenize public attitudes at a scale without employing overt disinformation (Agarwal et al., 2025; Daryani et al., 2026).

The widespread proliferation of AI-generated content driving this homogenization of discourse appears to be largely driven by economic incentives rather than coordinated intent (Santy et al., 2025; Zhang and Zhang, 2024). However, this crowding out of human involvement in the production of text on the internet exposes a vulnerability in existing platform governance: while online platforms possess robust infrastructures to detect and moderate overt harms, such as hate speech (Gorwa et al., 2020; Hee et al., 2024) or, to some extent, factual inaccuracies (Tokita et al., 2024; Westlund et al., 2024), they are unequipped to govern for semantic diversity or epistemic quality (Gorwa et al., 2020; Lasser and Poechhacker, 2025; Palla et al., 2025).

In addition to impacts on online discourse, this shift in the makeup of the internet carries immediate consequences for AI research and development. The 35% prevalence of AI-generated and AI-assisted text transforms the theoretical risk of model collapse (Schaeffer et al., 2025; Shumailov et al., 2024), wherein future AI models degrade after recursively ingesting AI-generated data, into an empirical concern. Future foundation models trained on contemporary data crawled from the internet will inevitably ingest a dataset that is, to a large extent, AI-generated and substantially less semantically diverse. While this may have a practical impact on pre-training, post-training and alignment stages will likely not be impacted, as they mostly utilize newly generated or environment-based data (Dohmatob et al., 2024b; Gan and Liu, 2024). It also remains an open question whether these recursive degradations are triggered by self-ingestion within a single model lineage or if they persist across a heterogeneous ecosystem of disparate models (Alemohammad et al., 2023; Gerstgrasser et al., 2024).

While our findings introduce concerns for productive democratic discourse on the internet, AI-generated text online should not be understood as inherently negative or as having intrinsic moral value. We see many scenarios in which it may democratize access to online conversation, such as empowering non-native speakers and individuals with varying literacy levels to participate more fluently in the online public sphere (Baldrich et al., 2025; Kalantzis and Cope, 2025; Tafazoli, 2024). Other positive use cases might include automated summarization of complicated documents for information accessibility (Marturi and Elwazzan, 2025), and the mass-scale localization of global

knowledge for underserved, low-resource languages (Ankinina et al., 2025; Merx et al., 2024).

Countering the unintended consequences of AI proliferation on the internet requires acknowledging the fundamental limits of post-hoc detection. While the detection of AI-generated text remains reliable for now (see Appendix A), in some contexts it is intractable (e.g., when the text is very short), and the confidence of these methods may change over time. Mandates relying on retroactive detection or easily circumvented text watermarking (Cheng et al., 2025; Migliorini, 2024; Nemecek et al., 2025; Rijsbosch et al., 2025), which have already been passed in numerous countries (European Commission, 2024; European Union, 2024), may therefore be inadequate. Preserving an open discourse with verifiable human participation may instead require a pivot toward cryptographically verifying human provenance (e.g., C2PA-like standards) (C2PA, 2024; Kaye and Dixon, 2025) and recalibrating search and recommendation algorithms to reward semantic diversity and verified human origin over raw content volume or engagement (Lasser and Poechhacker, 2025; Yu et al., 2024).

Even though factual accuracy may not be a concern in online discourse at present, future research should examine whether factual accuracy degrades as recursive training feedback loops accelerate across subsequent model generations. Additionally, while this study offers a baseline for text, analogous web-scale prevalence studies are urgently needed for multimodal content (Chandra et al., 2025; Croitoru et al., 2024). In the context of global elections, researchers should carefully distinguish between the ambient proliferation of financially motivated AI content and the targeted intent of adversarial campaigns, as human intent remains an essential variable in protecting democratic discourse.

It is important to note that the impact of AI-generated imagery may be fundamentally different from text; while text-based synthetic proliferation primarily causes semantic contraction, deepfake imagery poses a more direct threat to visual evidence and can trigger more visceral forms of systemic truth decay due to the historical trust placed in photographic documentation.

3. Methods

3.1. Data Collection

To obtain a representative sample of websites published on the internet, we use the longitudinal URL sampling methodology introduced by Garg et al. (2025) to draw samples from the Wayback Machine at the Internet Archive.[1] This approach employs multi-dimensional stratified sampling of the Internet Archive's CDX index, which catalogs every archived web page. Specifically, URLs are sampled along several dimensions—time of first archival, MIME type, URL depth, and top-level domain—to mitigate biases arising from the Archive's increasing crawl capacity over time and the over-representation of popular domains. Logarithmic-scale downsampling is applied to reduce the influence of highly archived domains, and top-level URLs are extracted from deep links to upsample earlier periods. The resulting sample is designed to approximate a uniform random draw from the population of publicly accessible web pages archived during the sampling period.

We sample websites from the Internet Archive spanning 33 monthly intervals from August 2022 to May 2025. For each sampled URL, we retrieve the oldest available archived snapshot via the Wayback Machine's CDX Server API.[2] The raw HTML of each snapshot is downloaded and stored locally for subsequent processing.

From each archived HTML page, we extract the visible text—i.e., the textual content that would be rendered and visible to a user visiting the page in a web browser. We accomplish this using the Trafilatura (Barbaresi, 2021) library, which is designed to isolate the main textual content of webpages while removing boilerplate elements such as navigation menus, headers, footers, and other non-content components. The extracted content is segmented into paragraphs, and the longest paragraph (measured by word count) is selected as the representative text sample for that page. If the resulting paragraph contains fewer than 100 words, the document is excluded from further

[1] We note any modifications on top of the sampling methodology used by Garg et al. (2025) in Appendix D.

[2] See https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server.

analysis. Because interface elements injected by the Wayback Machine (e.g., replay banners or error messages) are typically short, these steps make it unlikely that such artifacts appear in the final dataset. In addition, language detection is applied using the `langdetect` (Danilak, 2014) library, and only English text is retained.

3.2. Participant Study

To assess public beliefs about the impact of AI-generated text on the internet (RQ1), we conducted a survey of adults in the United States, recruited through Prolific. The sample was stratified to be representative of the US adult population with respect to age, sex, and ethnicity.

The study was administered as an online form in three parts, each corresponding to a subset of the six hypotheses. Part 1 addressed Hypotheses 1–3 (n = 303), Part 2 addressed Hypotheses 4 and 6 (n = 301), and Part 3 addressed Hypothesis 5 (n = 299). In total, N = 903 responses were collected from 853 unique participants. Each part presented participants with a hypothesis statement and asked them to indicate their level of agreement on a 7-point Likert scale (Strongly Disagree, Disagree, Somewhat Disagree, Neither Agree nor Disagree, Somewhat Agree, Agree, Strongly Agree). In addition, each part collected two covariates: frequency of AI tool usage (Never, Monthly, Weekly, Daily) and general view of AI's impact on society (7-point scale from Very Negative to Very Positive).

The combined sample had a mean age of 45.4 years (SD = 15.73, range: 18–84) and was 50.9% female and 48.6% male. The ethnic composition was 62.6% White, 12.0% Black, 11.1% Mixed, 7.6% Other, and 6.3% Asian. Nearly all participants (99.6%) resided in the United States. Regarding AI tool usage, 39.8% reported daily use, 31.3% weekly, 17.1% monthly, and 11.2% never. In terms of general views of AI's impact on society, 46.0% held a positive view (Somewhat Positive, Positive, or Very Positive), 40.7% held a negative view (Somewhat Negative, Negative, or Very Negative), and 12.6% were neutral.

3.3. AI-Generated Text Detection

Detecting AI-generated text is a prerequisite for all subsequent analyses. We evaluate four detectors: Binoculars (Hans et al., 2024), Desklib (Desklib, 2025), DivEye (Basani and Chen, 2025), and the Pangram v3 commercial API (Pangram Labs, 2026). These were chosen based on their RAID benchmark performance (Dugan et al., 2024) and availability. While AI text detection is known to have its limitations, we believe it provides the best estimate of AI text prevalence in in our empirical setting (i.e., a collection of texts without available watermarking). We create a custom set of tests to better understand these limitations for our use case, and choose the most suitable detector, as reported below.

We compare the four above-mentioned detectors across five dimensions: (1) text length sensitivity, testing detection performance on texts ranging from 1 to 500 words; (2) HTML robustness, comparing detection accuracy on identical AI-generated text in plain versus HTML-embedded formats; (3) model family, evaluating detection of outputs from GPT-4o, Claude, and Gemini; (4) model version, testing across OpenAI model versions from davinci-002 to GPT-4o; and (5) multilingual robustness. The full results of this robustness analysis are reported in Appendix A.

Based on this evaluation, we selected the Pangram v3 API for our primary analyses. Pangram v3 outperformed the three remaining methods across most robustness dimensions, achieving perfect accuracy on texts exceeding 50 words across languages and in both raw text and HTML-wrapped formats. Binoculars and DivEye proved particularly unreliable: Binoculars exhibited an 11.4 percentage point drop in detection rate when AI-generated text was embedded in HTML, and showed substantially higher difficulty detecting Claude-generated text; DivEye showed no overlap between plain text and HTML-embedded score distributions and failed to separate AI-generated from human-written text across most non-English languages. Desklib performed comparably to Pangram v3 on text length and model family robustness, and even outperformed it on model version robustness; however, it underperformed

on HTML versus plain text and language robustness, which we consider more critical given that our corpus consists of multilingual archived web pages. In contrast, Pangram v3 maintained stable performance across all five evaluated dimensions. An additional advantage of Pangram v3 over the three other detectors is that it operates in a three-way classification scheme — AI-generated, AI-assisted, and human-written — that provides richer signal than binary detection.

Pangram v3 classifies each input text into three categories: (1) fully AI-generated (`fraction_ai`), (2) AI-assisted, i.e., human-written with AI involvement (`fraction_ai_assisted`), and (3) fully human-written (`fraction_human`), where the three fractions sum to approximately one. The detector operates over sliding windows of the input text and aggregates segment-level predictions into document-level scores.

For each website in our sample, we report three scores corresponding to these categories. We define the *aggregate AI likelihood score* for a given monthly sample as a combined metric incorporating both the fully AI-generated and AI-assisted fractions, which serves as the primary independent variable in our hypothesis tests. In the results, we additionally report the proportion of websites classified as fully AI-generated (i.e., where `fraction_ai` exceeds a threshold) and the proportion classified as either AI-generated or AI-assisted.

3.4. Hypothesis Verification

We test six hypotheses about the impact of AI-generated text on the internet. For each hypothesis, we define a measurable signal, compute it for each monthly sample of websites, and test whether it correlates with the aggregate AI likelihood score across months. All correlations are evaluated using the Pearson correlation coefficient; we reject the null hypothesis (H_0: $\rho = 0$) at the $\alpha = 0.05$ significance level.

Hypothesis 1: Semantic Contraction. We operationalize semantic diversity as the average pairwise cosine similarity of document-level semantic embeddings within each monthly sample. Embeddings are computed using the `all-MiniLM-L6-v2` sentence embedding model (Reimers and Gurevych, 2019) from the `sentence-transformers` library[3]. Each document's extracted visible text is encoded into a 384-dimensional dense vector. For samples with $n \leq 500$ documents, we compute all $\binom{n}{2}$ pairwise cosine similarities using `scipy.spatial.distance.pdist`; for larger samples, we estimate the mean pairwise similarity by randomly sampling up to 10,000 document pairs. Embeddings are L2-normalized prior to similarity computation. A positive correlation between this signal and the aggregate AI likelihood score would indicate semantic contraction.

Hypothesis 2: Truth Decay. We measure factual accuracy using a two-stage pipeline: automated claim extraction followed by human fact-checking. In the first stage, we use `GPT-4o-mini` (Hurst et al., 2024) via the OpenAI API to extract up to five verifiable factual claims from each website's visible text. The extraction prompt instructs the model to identify specific, self-contained assertions of fact (e.g., statistics, dates, named entities) while excluding opinions, predictions, and vague statements. Input text is truncated to 6,000 characters and the model is queried at temperature 0.1 to maximize determinism.

In the second stage, extracted claims are verified by human annotators recruited on Prolific. Annotators use a custom web-based annotation interface (see Appendix B) to assess each claim against one of four verdict categories, following a scheme inspired by FEVER (Thorne et al., 2018) and AVeriTeC (Schlichtkrull et al., 2024): *Supported* (claim is corroborated by reliable evidence), *Refuted* (claim is contradicted by reliable evidence), *Not Enough Evidence* (insufficient evidence to verify), and *Conflicting Evidence* (sources both support and refute the claim). Annotators are instructed to search for evidence using established sources (e.g., websites of major news, governmental, and academic organizations) and to record the URLs of all consulted sources. Each

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

annotator evaluates claims from five articles; a 20% overlap in claim assignments is maintained to compute the inter-annotator agreement.

The fact-checking study recruited $N = 50$ approved annotators (after excluding 11 timed-out submissions), with a mean age of 39.8 years (SD = 12.39, range: 21–70), 62.0% male and 38.0% female. Annotators were located across multiple countries (36.0% United Kingdom, 18.0% United States, 12.0% Portugal, and others). The mean annotation time was 41.3 minutes (SD = 16.1). We compute inter-annotator agreement using Krippendorff's alpha (Krippendorff, 2004), which accommodates missing data from incomplete overlap assignments.

We define the factual error rate as the proportion of claims labeled as *Refuted* within each monthly sample. A positive correlation between this rate and the aggregate AI likelihood score would indicate truth decay.

Hypothesis 3: Positivity Shift. We measure document-level sentiment using the `cardiffnlp/twitter-roberta-base-sentiment-latest` model (Loureiro et al., 2022), a `RoBERTa-base` model fine-tuned on 124M tweets for sentiment classification[4]. Each document's visible text is truncated to 500 characters and tokenized with a maximum sequence length of 512 tokens. The model outputs softmax probabilities over three classes (negative, neutral, positive), and each document is assigned the label with the highest probability. We compute the rate of documents classified as positive within each monthly sample. A positive correlation between this rate and the aggregate AI likelihood score would indicate a positivity shift.

Hypothesis 4: Epistemic Islands. We measure the density of outbound hyperlinks as a proxy for epistemic connectivity. For each archived HTML page, we parse the DOM using BeautifulSoup with the `lxml` parser. We first remove structural navigation elements (`<nav>`, `<footer>`, `<header>`, `<aside>`) to isolate the article body, then count all remaining anchor tags (`<a>`) with `href` attributes. Link density is computed as the number of outbound links per 1,000 words of visible text; documents with fewer than 50 words are excluded. We compute the mean link density for each monthly sample. An inverse correlation between link density and the aggregate AI likelihood score would indicate epistemic island formation.

Hypothesis 5: Entropy Dilution. We measure the information density of documents using the Gzip compression ratio. For each document, we encode the extracted visible text as UTF-8, compute the size of the raw byte string, compress it using the `gzip` module from the Python standard library, and compute the ratio of compressed size to raw size. Lower compression ratios indicate higher information density (i.e., less redundancy), while higher ratios indicate more compressible, repetitive content. Documents with fewer than 100 characters of extracted text are excluded. A positive correlation between the mean compression ratio and the aggregate AI likelihood score would indicate entropy dilution.

Hypothesis 6: Stylistic Monoculture. We operationalize stylistic diversity using character-level n-gram similarity. For each document, we extract all character 3-grams from the first 2,000 characters of the lowercased visible text and represent each document as a set of unique 3-grams. Pairwise stylistic similarity between documents is computed using the Jaccard index (i.e., the ratio of the intersection to the union of the two 3-gram sets). For samples with $n \leq 100$ documents, we compute all pairwise similarities; for larger samples, we randomly draw up to 10,000 pairs. The mean Jaccard similarity for each monthly sample serves as the signal. A positive correlation between this signal and the aggregate AI likelihood score would indicate stylistic monoculture.

Statistical Testing. For all six hypotheses, the independent variable is the aggregate AI likelihood score (combining Pangram v3's fully AI-generated and AI-assisted fractions) aggregated at the monthly level, and the dependent variable

[4] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

is the corresponding monthly signal. We compute the Pearson correlation coefficient ρ and its associated p-value. Significance is assessed at $\alpha = 0.05$.

Acknowledgements

The authors thank the Internet Archive for providing the Wayback Machine data and their extraordinary support throughout this project and Pangram for providing a research grant supporting the use of their API. We also thank Liam Dugan, Hany Farid, Daphne Ippolito, Shayne Longpre, and Alexander Wang for helpful discussions (listed in alphabetical order).

Competing Interests

Pangram had no role in study design, data collection, analysis, interpretation, or the decision to publish. We only applied for the research grant after conducting the robustness analysis and deciding to use their detector.

References

D. Agarwal, M. Naaman, and A. Vashistha. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–21, 2025.

S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. Baraniuk. Self-consuming generative models go MAD. *arXiv preprint arXiv:2307.01850*, 2023.

S. Altay and F. Gilardi. People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS nexus*, 3(10):pgae403, 2024.

T. Ankinina, J. Cegin, J. Simko, and S. Ostermann. A rigorous evaluation of LLM data generation strategies for low-resource languages. In *Conference on Empirical Methods in Natural Language Processing*, 2025.

K. Baldrich, C. Pérez-García, and M. Santamarina-Sancho. Artificial intelligence in academic literacy: Empirical evidence on reading and writing practices in higher education. In *Frontiers in Education*, volume 10, page 1701238. Frontiers Media SA, 2025.

A. Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the ACL 2021 System Demonstrations*, pages 122–131, 2021.

A. R. Basani and P.-Y. Chen. Diversity boosts AI-generated text detection. *arXiv preprint arXiv:2509.18880*, 2025.

C2PA. Content credentials — C2PA technical specification, 2024. URL https://spec.c2pa.org/specifications/specifications/2.1/specs/_attachments/C2PA_Specification.pdf.

N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, et al. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

J. Chein, S. Martinez, and A. Barone. Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from AI texts. *Scientific Reports*, 14(1):25989, 2024.

S. Chen, M. Gao, K. Sasse, T. Hartvigsen, B. Anthony, L. Fan, H. Aerts, J. Gallifant, and D. S. Bitterman. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 8(1):605, 2025.

Y. Cheng, H. Guo, Y. Li, and L. Sigal. Revealing weaknesses in text watermarking through self-information rewrite attacks. *arXiv preprint arXiv:2505.05190*, 2025.

B. Chesney and D. Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea, P. Irofti, M. Popescu, C. Rusu, R. T. Ionescu,

F. S. Khan, and M. Shah. Deepfake media generation and detection in the generative AI era: A survey and outlook. *arXiv preprint arXiv:2411.19537*, 2024.

M. Danilak. langdetect: Language detection library (python port). https://github.com/Mimino666/langdetect, 2014.

S. Daniotti, J. Wachs, X. Feng, and F. Neffke. Who is using AI to code? global diffusion and impact of generative AI. *Science*, page eadz9311, 2026.

Y. Daryani, Z. Sourati, and M. Dehghani. The homogenizing engine: AI's role in standardizing culture and the path to policy. *Policy Insights from the Behavioral and Brain Sciences*, 13(1): 14–27, 2026.

H. Dawkins, K. C. Fraser, and S. Kiritchenko. When detection fails: The power of fine-tuned models to generate human-like social media text. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13494–13527, 2025.

Desklib. AI text detector v1.01, 2025. URL https://huggingface.co/desklib/ai-text-detector-v1.01.

E. Dohmatob, Y. Feng, and J. Kempe. Model collapse demystified: The case of regression. *Advances in Neural Information Processing Systems*, 37:46979–47013, 2024a.

E. Dohmatob, Y. Feng, A. Subramonian, and J. Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024b.

L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, and C. Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, 2024.

European Commission. Regulatory framework for AI, 2024. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

European Union. Article 50 — artificial intelligence act, 2024. URL https://artificialintelligenceact.eu/article/50/.

E. Ferrara. The generative AI paradox: GenAI and the erosion of trust, the corrosion of information verification, and the demise of truth. *Future Internet*, 18(2):73, 2026.

K. C. Fraser, H. Dawkins, and S. Kiritchenko. Detecting AI-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, 82:2233–2278, 2025.

Z. Gan and Y. Liu. Towards a theoretical understanding of synthetic data in LLM post-training: A reverse-bottleneck perspective. *arXiv preprint arXiv:2410.01720*, 2024.

K. Garg, S. Alam, D. Ayala, M. Graham, M. C. Weigle, and M. L. Nelson. Longitudinal sampling of URLs from the wayback machine. *arXiv preprint arXiv:2507.14752*, 2025.

M. Gerstgrasser, R. Schaeffer, A. Dey, R. Rafailov, H. Sleight, J. Hughes, T. Korbak, R. Agrawal, D. Pai, A. Gromov, D. A. Roberts, D. Yang, D. L. Donoho, and O. Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945, 2020.

A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024.

M. S. Hee, S. Sharma, R. Cao, P. Nandi, P. Nakov, T. Chakraborty, and R. K.-W. Lee. Recent advances in online hate speech moderation: Multimodality and the role of large models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419, 2024.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

C. Jacob, P. Kerrigan, and M. Bastos. The chat-chamber effect: Trusting the AI hallucination. *Big data & society*, 12(1):20539517241306345, 2025.

M. Jakesch, J. T. Hancock, and M. Naaman. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023.

M. Kalantzis and B. Cope. Literacy in the time of artificial intelligence. *Reading Research Quarterly*, 60(1):e591, 2025.

K. Kaye and P. Dixon. Privacy, identity and trust in C2PA: A technical review and analysis of the C2PA digital media provenance framework, 2025. URL https://worldprivacyforum.org/posts/privacy-identity-and-trust-in-c2pa/.

B. Kitchens, S. L. Johnson, and P. Gray. Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, 44(4):1619–1649, 2020.

D. Kobak, R. González-Márquez, E.-Á. Horvát, and J. Lause. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025.

K. Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30 (3):411–433, 2004.

L. La Cava, L. M. Aiello, and A. Tagarelli. Machines in the crowd? measuring the footprint of machine-generated text on reddit. *arXiv preprint arXiv:2510.07226*, 2025.

J. Lasser and N. Poechhacker. Designing social media content recommendation algorithms for societal good. *Annals of the New York Academy of Sciences*, 1548(1):20–28, 2025.

W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, L. Chen, H. Ye, S. Liu, Z. Huang, et al. Monitoring AI-modified content at scale: A case study on the impact of chatgpt on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024.

Q. Liu, L. Wang, and M. Luo. When seeing is not believing: self-efficacy and cynicism in the era of intelligent media. *Humanities and Social Sciences Communications*, 12(1):1–13, 2025.

D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados. TimeLMs: Diachronic language models from twitter. In *Proceedings of the 60th annual meeting of the association for computational linguistics: System demonstrations*, pages 251–260, 2022.

L. Malmqvist. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 61–74. Springer, 2025.

K. C. Marturi and H. H. Elwazzan. LLM-guided planning and summary-based scientific text simplification: DS@GT at CLEF 2025 SimpleText. *arXiv preprint arXiv:2508.11816*, 2025.

H. Matatov, M. A. L. Quéré, O. Amir, and M. Naaman. Examining the prevalence and dynamics of AI-generated media in art subreddits. *arXiv preprint arXiv:2410.07302*, 2024.

R. Merx, H. Suominen, A. J. G. Correia, T. Cohn, and E. Vylomova. Low-resource machine translation: what for? who for? an observational study on a dedicated tetun language translation service. *arXiv preprint arXiv:2411.12262*, 2024.

S. Migliorini. China's interim measures on generative AI: Origin, content and significance. *Computer Law & Security Review*, 53:105985, 2024.

C. Mouffe. Deliberative democracy or agonistic pluralism? In *Models of Deliberative Democracy*, pages 251–264. Routledge, London, 1999.

P. Muzumdar, S. Cheemalapati, S. R. RamiReddy, K. Singh, G. Kurian, and A. Muley. The dead internet theory: a survey on artificial interactions and the future of social media. *arXiv preprint arXiv:2502.00007*, 2025.

A. Nemecek, Y. Jiang, and E. Ayday. Watermarking without standards is not AI governance. *arXiv preprint arXiv:2505.23814*, 2025.

D. Oh and J. Downey. Does algorithmic content moderation promote democratic discourse? radical democratic critique of toxic language AI. *Information, Communication & Society*, 28 (7):1157–1176, 2025.

K. Palla, J. L. R. García, C. Hauff, F. Fabbri, A. Damianou, H. Lindström, D. Taber, and M. Lalmas. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 840–854, 2025.

Pangram Labs. Pangram v3 AI content detection API. https://www.pangram.com/solutions/api, 2026.

J. L. Paredes, E. Smith, G. Druck, and B. Benson. More articles are now created by AI than humans, 2025. URL https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans.

B. Porter and E. Machery. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133, 2024.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

B. Rijsbosch, G. van Dijck, and K. Kollnig. Adoption of watermarking for generative AI systems in practice and implications under the new eu AI act. *arXiv preprint arXiv:2503.18156*, 2025.

J. Russell, M. Karpinska, D. Akinode, K. Thai, B. Emi, M. Spero, and M. Iyyer. AI use in american newspapers is widespread, uneven, and rarely disclosed. *arXiv preprint arXiv:2510.18774*, 2025.

V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025.

K. Saito, A. Wachi, K. Wataoka, and Y. Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

S. Santy, P. Bhattacharya, M. H. Ribeiro, K. Allen, and S. Oh. When incentives backfire, data stops being human. *arXiv preprint arXiv:2502.07732*, 2025.

R. Schaeffer, J. Kazdan, A. C. Arulandu, and S. Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.

K. J. Schiff, D. S. Schiff, and N. S. Bueno. The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, 119(1):71–90, 2025.

M. Schlichtkrull, Y. Chen, C. Whitehouse, Z. Deng, M. Akhtar, R. Aly, Z. Guo, C. Christodoulopoulos, O. Cocarascu, A. Mittal, et al. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, 2024.

M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse

when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

D. H. Spennemann. Delving into: the quantification of AI-generated content on the internet (synthetic data). *arXiv preprint arXiv:2504.08755*, 2025.

Z. Sun, Z. Zhang, X. Shen, Z. Zhang, Y. Liu, M. Backes, Y. Zhang, and X. He. Are we in the AI-generated text world already? quantifying and monitoring AIGT on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005, 2025.

D. Tafazoli. Exploring the potential of generative AI in democratizing english language education. *Computers and Education: Artificial Intelligence*, 7:100275, 2024.

B. Thompson, M. Dhaliwal, P. Frisch, T. Domhan, and M. Federico. A shocking amount of the web is machine translated: Insights from multiway parallelism. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1763–1775, 2024.

H. S. Thompson. Improved methodology for longitudinal web analytics using common crawl. In *Proceedings of the 16th ACM Web Science Conference*, pages 59–69, 2024.

J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

C. K. Tokita, K. Aslett, W. P. Godel, Z. Sanderson, J. A. Tucker, J. Nagler, N. Persily, and R. Bonneau. Measuring receptivity to misinformation at scale on a social media platform. *PNAS nexus*, 3(10):pgae396, 2024.

O. Westlund, V. Belair-Gagnon, L. Graves, R. Larsen, and S. Steensen. What is the problem with misinformation? fact-checking as a sociotechnical and problem-solving practice. *Journalism Studies*, 25(8):898–918, 2024.

J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1): 275–338, 2025.

S. Xing, J. Hong, Y. Wang, R. Chen, Z. Zhang, A. Grama, Z. Tu, and Z. Wang. Llms can get" brain rot"! *arXiv preprint arXiv:2510.13928*, 2025.

X. Yu, M. Haroon, E. Menchen-Trevino, and M. Wojcieszak. Nudging recommendation algorithms increases news consumption and diversity on YouTube. *PNAS nexus*, 3(12):pgae518, 2024.

J. Zhang, S. Yu, D. Chong, A. Sicilia, M. R. Tomz, C. D. Manning, and W. Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*, 2025.

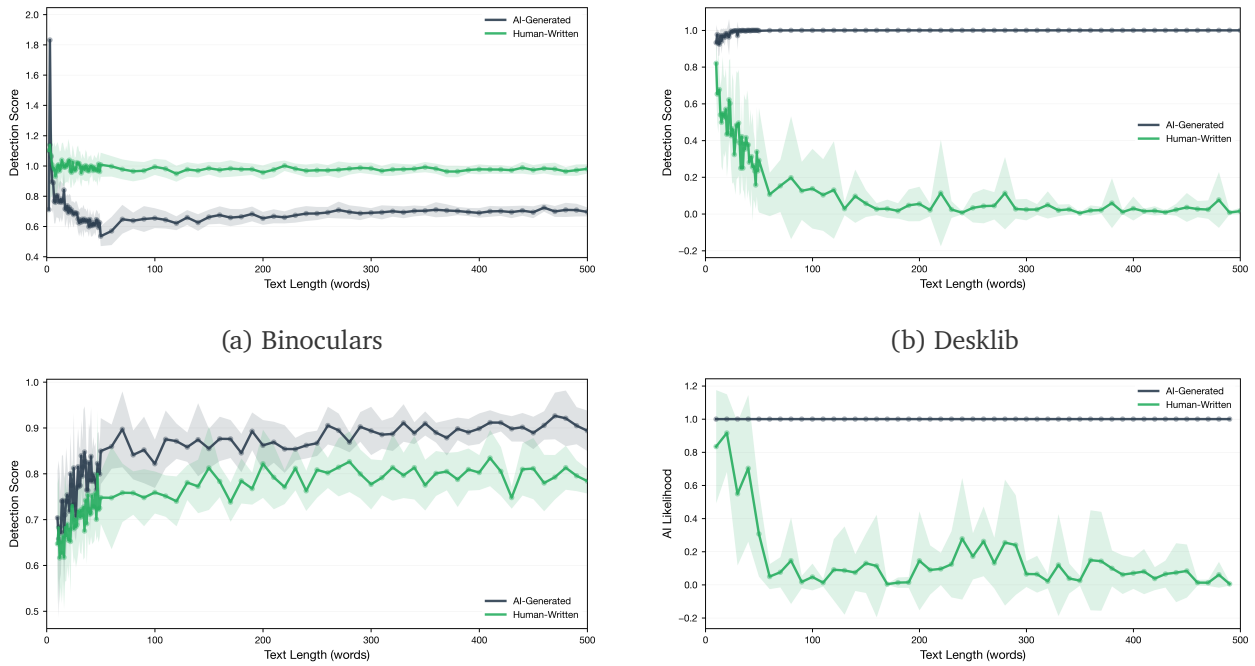
Y. Zhang and T. Zhang. The impact of generative AI on content platforms: A two-sided market analysis with multi-dimensional quality heterogeneity. *arXiv preprint arXiv:2410.13101*, 2024.

A. Robustness Analysis of AI-Generated Text Detectors

We conducted a systematic robustness analysis comparing four detectors of AI-generated text—Binoculars (Hans et al., 2024), Desklib (Desklib, 2025), DivEye (Basani and Chen, 2025), and the Pangram v3 commercial API (Pangram Labs, 2026)—across five experimental dimensions to inform our choice of detector for the main analyses.

Text Length. We evaluated detection robustness across texts of various lengths, ranging from 1 to 500 words, using 20 AI-generated and 20 human-written samples per length bin. Length sensitivity curves for all detectors are shown in Figure 4, indicating the separability between AI-generated and human-written texts.

For texts of 1–10 words, Binoculars achieved a true positive rate (TPR) of 61.6% with a false positive rate (FPR) of 35%, rendering it unreliable for very short texts. DivEye also achieved limited separability with shorter texts. Meanwhile, Desklib and Pangram v3 scores between the distributions were smaller but still maintained separability. Performance improved substantially for texts of 11–50 words (for Binoculars: TPR: 96.6% and FPR: 17.9%). Binoculars, Desklib, and Pangram v3 reached near-perfect levels for texts exceeding 50 words. DivEye improved in these ranges, but registered less separability.



(a) Binoculars

(b) Desklib

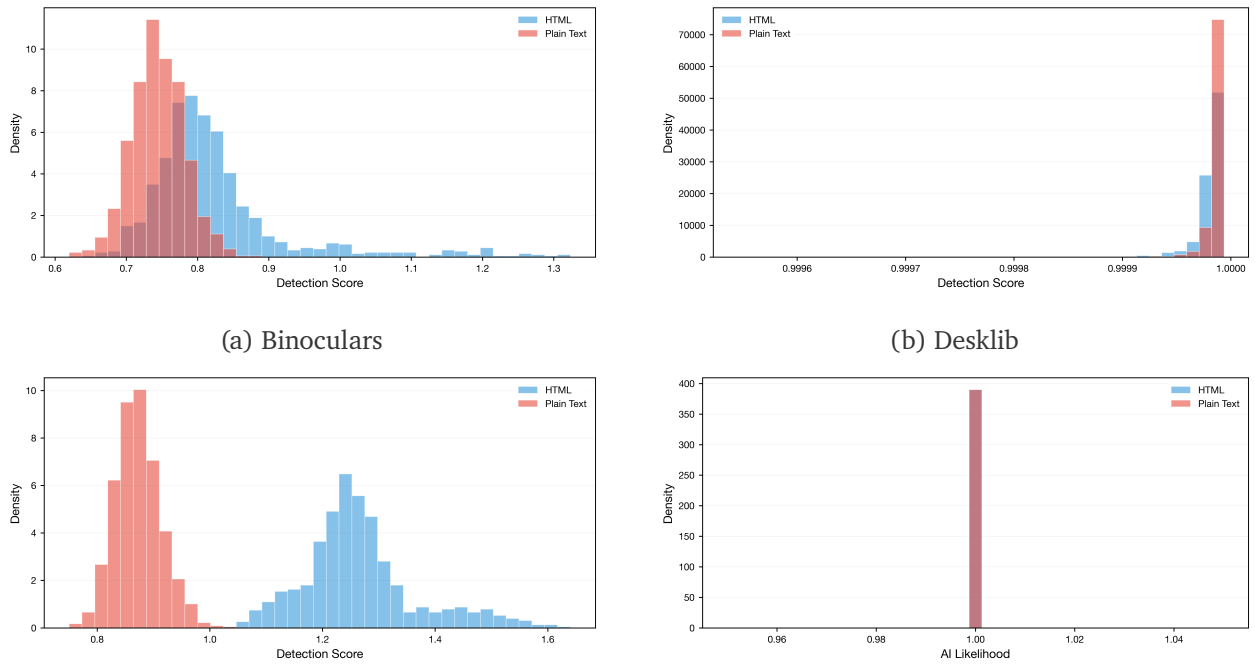
(c) DivEye

(d) Pangram v3

Figure 4 | **Length Sensitivity Analysis.** Detection scores for (a) Binoculars, (b) Desklib, (c) DivEye, and (d) Pangram v3 across varying text lengths. Except for DivEye, these detectors generally achieve strong performance for texts exceeding 100 words.

HTML vs. Plain Text. We tested whether embedding AI-generated text within an HTML document structure affects detection accuracy. Identical AI-generated texts (produced by GPT-4o) were presented to all four detectors in both plain text and HTML-embedded formats. In Figure 5, we plot scores of these two distributions (plain text in red and HTML-embedded text in blue) for each tested model. For our use case, it would be ideal if the two distributions maintained a complete overlap, which would suggest that the distribution of AI-generated articles would generally register the same scores no matter whether wrapped in HTML or presented as plain text.

DivEye showed the largest divergence, wherein the two distributions had no overlap. Next, Binoculars showed partial divergence: the mean score shifted from 0.745 (100% flag rate, if detection threshold were applied) in plain text to 0.823 (88.6% flag rate, if detection threshold were applied) when HTML-embedded, an 11.4 percentage point drop in detection rate. Next, Desklib registered a minor shift between the two distributions; still, the scores for both distributions remained fully above 0.9999 (where 1.0 is the maximum possible score for AI-generated texts), and so this is not a major concern. Finally, Pangram v3 registered a complete perfect overlap between the two distributions.



(a) Binoculars

(b) Desklib

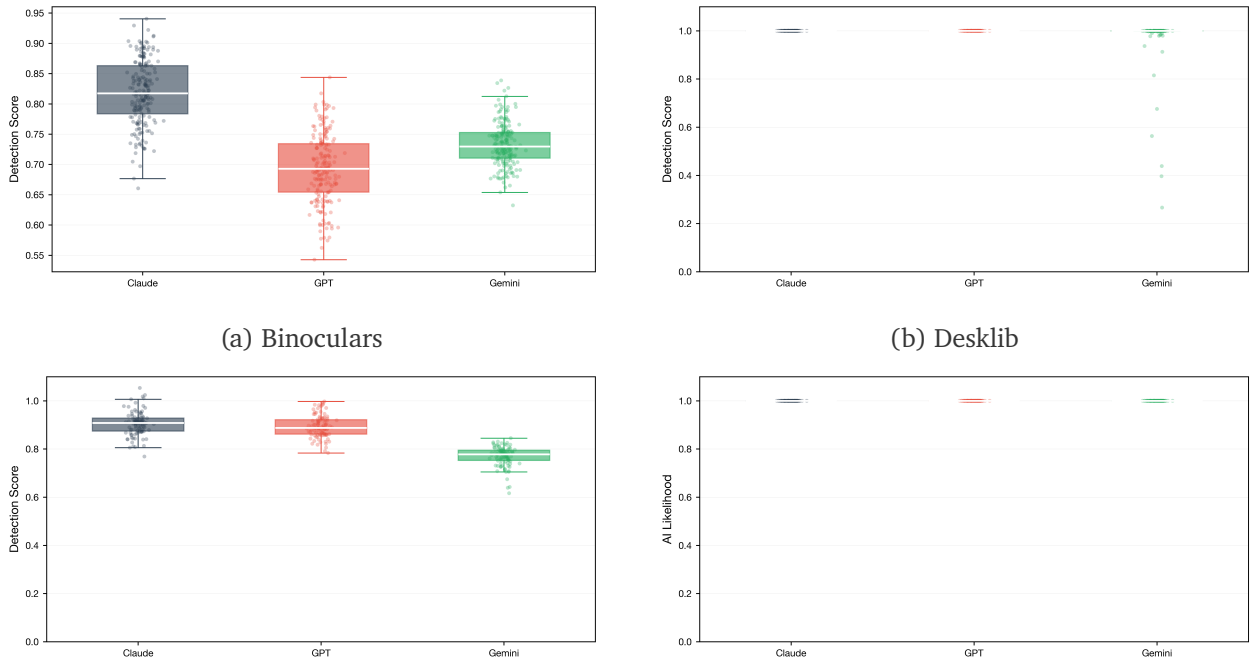
(c) DivEye

(d) Pangram v3

Figure 5 | **HTML vs. Plain Text Robustness.** Detection score distributions for AI-generated text presented in plain text versus HTML-embedded format, for (a) Binoculars, (b) Desklib, (c) DivEye, and (d) Pangram v3. Binoculars and DivEye exhibit a substantial drop in detection rate when text is embedded in HTML, Desklib exhibits some drop, and Pangram v3 exhibits minimal drop in scores.

Model Family. We evaluated model family robustness of the tested detectors across three leading LLM families: GPT-4o (OpenAI), Claude (Anthropic), and Gemini (Google DeepMind). Shown in Figure 6 are the distributions of scores for these model families registered by each of the tested detectors.

For Binoculars, GPT was easiest to detect, Gemini was second, and Claude was the hardest to detect. For DivEye, by contrast, Gemini was the hardest to detect, while Claude and GPT appeared similarly difficult. Compared to Claude and GPT, Desklib registered minor difficulty with Gemini in some cases, but generally performed similarly across the three. Pangram performed consistently across all three model families.



(a) Binoculars

(b) Desklib

(c) DivEye

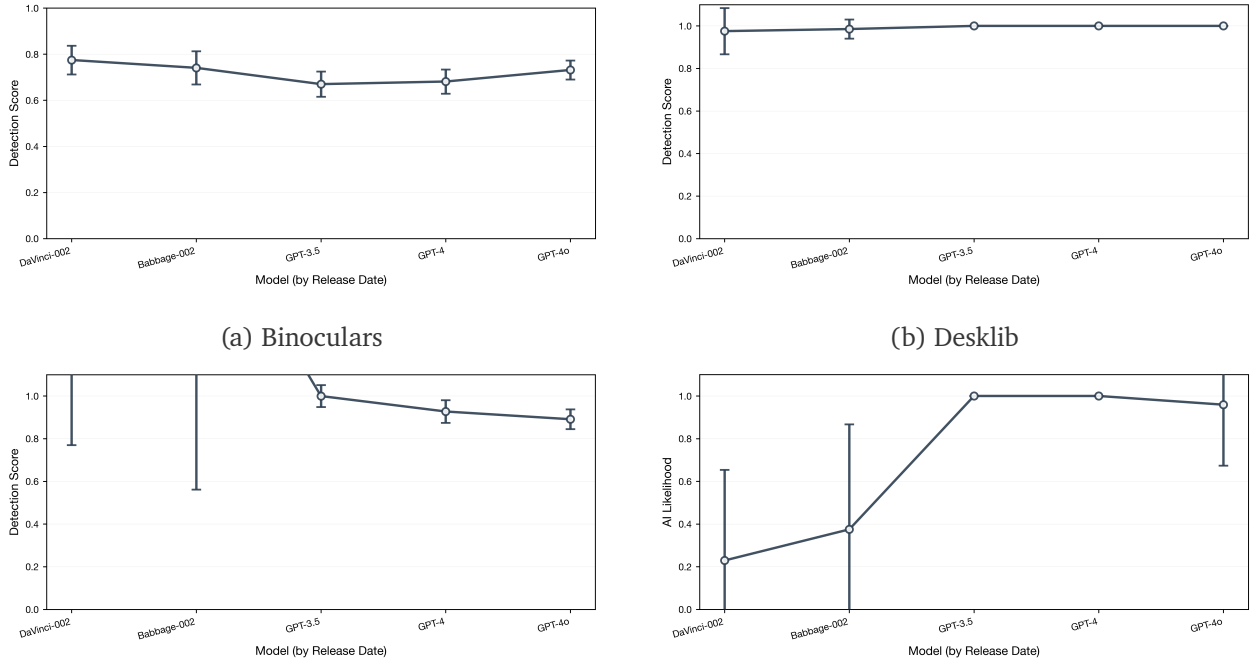
(d) Pangram v3

Figure 6 | **Model Family Robustness.** Detection score distributions for text generated by GPT-4o, Claude, and Gemini, for (a) Binoculars, (b) Desklib, (c) DivEye, and (d) Pangram v3. Binoculars performs differenly on Claude, GPT, and Gemini model families; Desklib and Pangram v3 have a similar performance across the board.

Model Version. We evaluated historical model version robustness across five generations of GPT-family models (OpenAI): `davinci-002`, `babbage-002`, `GPT-3.5`, `GPT-4`, and `GPT-4o`. The average scores, where the models are ordered from left to right by their date of release, are shown in Figure 7.

Binoculars and DivEye provide consistent results across all model versions. Pangram v3 does not detect texts generated by the two oldest model versions, `davinci-002` and `babbage-002`, which are old completion models. DivEye exhibited a similar inconsistency for completion-only models, but also registered a slightly decreasing performance from the later instruction-tuned `GPT-3.5` through `GPT-4o`.

While temporal consistency and robustness is generally important for best results, it is critical only for the instruction-tuned models which were available to the public through chatbot interfaces. The previous completion-only models were mostly used through APIs and generally discouraged for usage shortly after the public release of ChatGPT, and hence do not constitute a major concern.



(a) Binoculars

(b) Desklib

(c) DivEye

(d) Pangram v3

Figure 7 | **Model Version Robustness.** Detection scores across OpenAI model generations for (a) Binoculars, (b) Desklib, (c) DivEye, and (d) Pangram v3. Binoculars and Desklib perform consistently across the board; DivEye and Pangram v3 perform differently on completion models (`davinci-002` and `babbage-002`) compared to instruction-trained models (GPT-3.5, GPT-4, and GPT-4o).

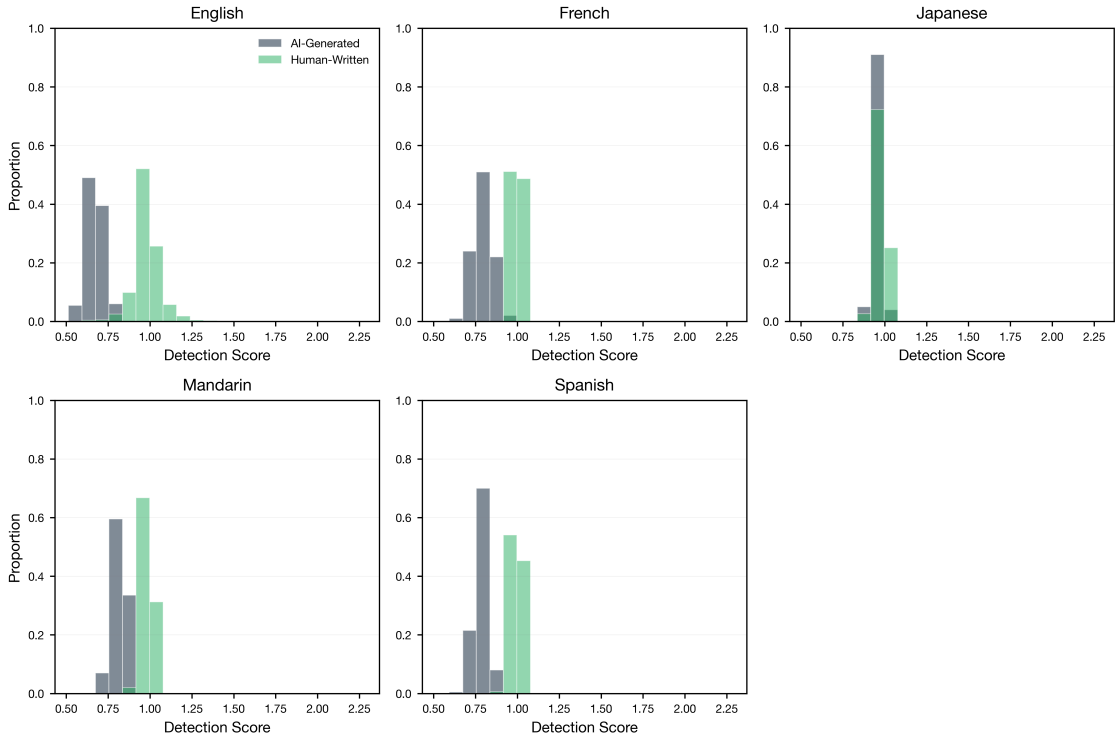
Language. We evaluated language robustness of the four tested detectors on AI-generated and human-written text across multiple languages: English, French, Japanese, Mandarin, and Spanish. Language-specific score distributions are shown in Figure 8.

Binoculars scores yielded nearly nonoverlapping distributions for English, French, Mandarin, and Spanish; however, the ranges of these distributions (and with them, the functional threshold) shifted. Japanese was not separable.

Desklib scores were bimodal and separable for English and Spanish. For French, Japanese, and Mandarin, the model behaved differently, with partial overlap.

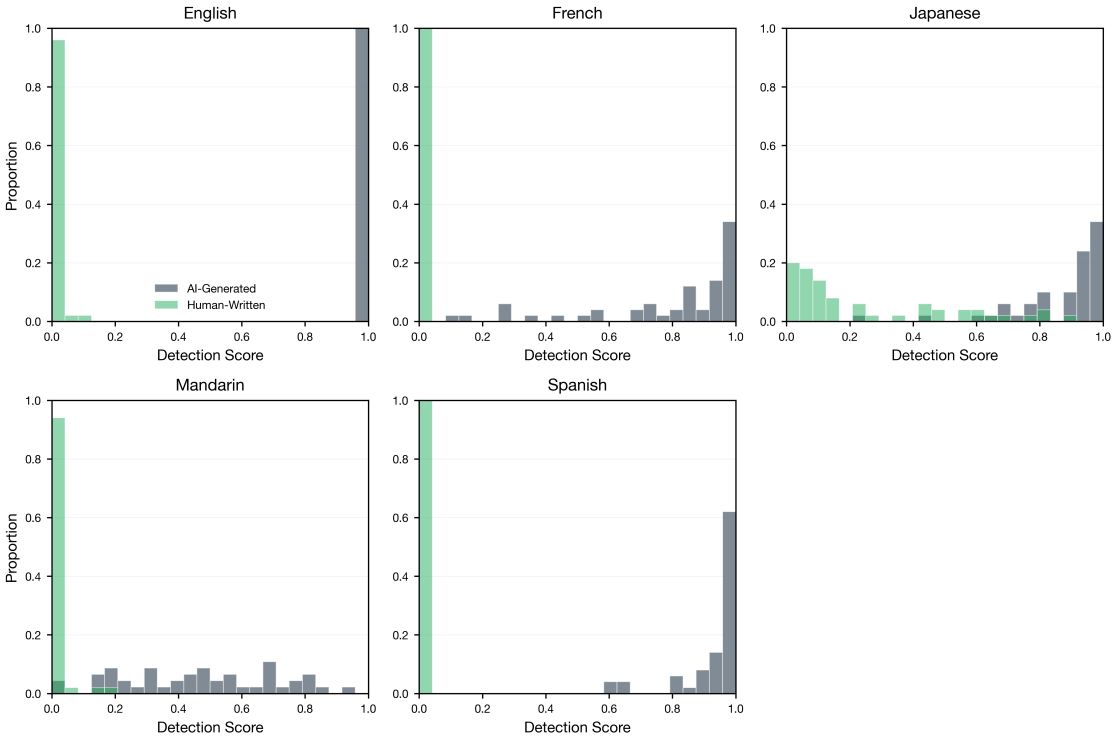
DivEye scores yielded nearly nonoverlapping distributions for English. For all other languages we tested (French, Japanese, Mandarin, and Spanish), the score distributions were largely not separable, and their ranges varied.

Pangram v3 scores yielded bimodal distributions for all tested languages. Notably, the detector behaved consistently regardless of the language of the input text.

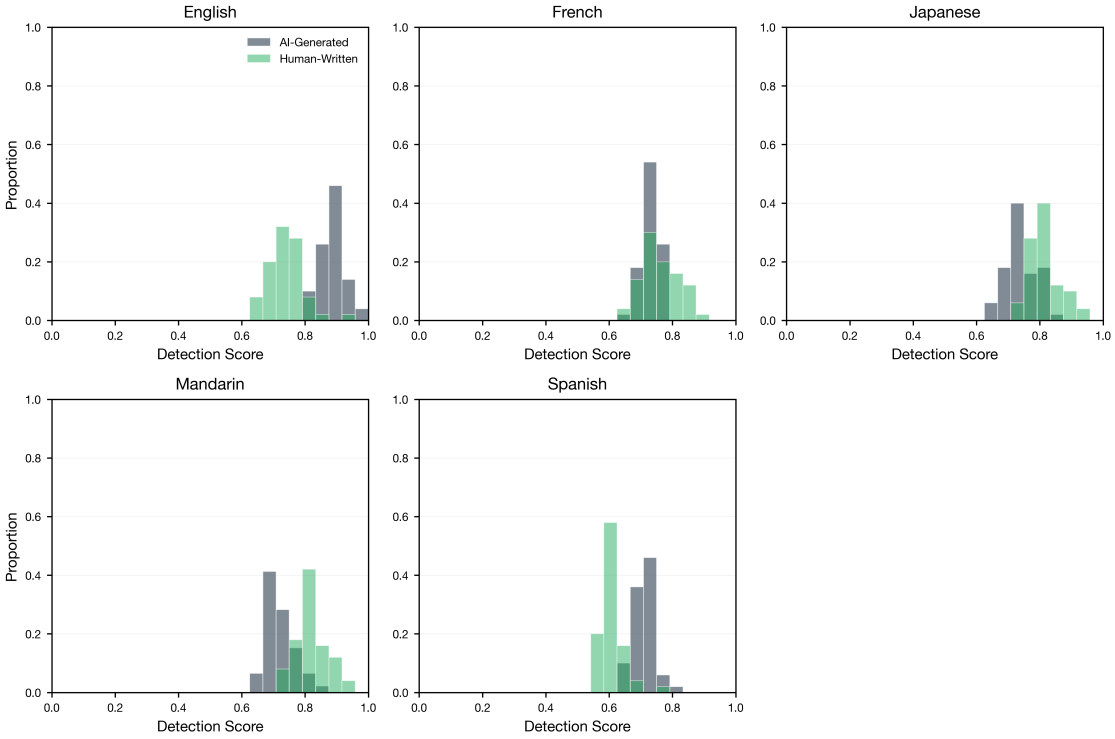


(a) Binoculars

Figure 8 | **Language Robustness.** Comparison of detector score distributions across languages for Binoculars, Desklib, DivEye, and Pangram v3 (1/3).

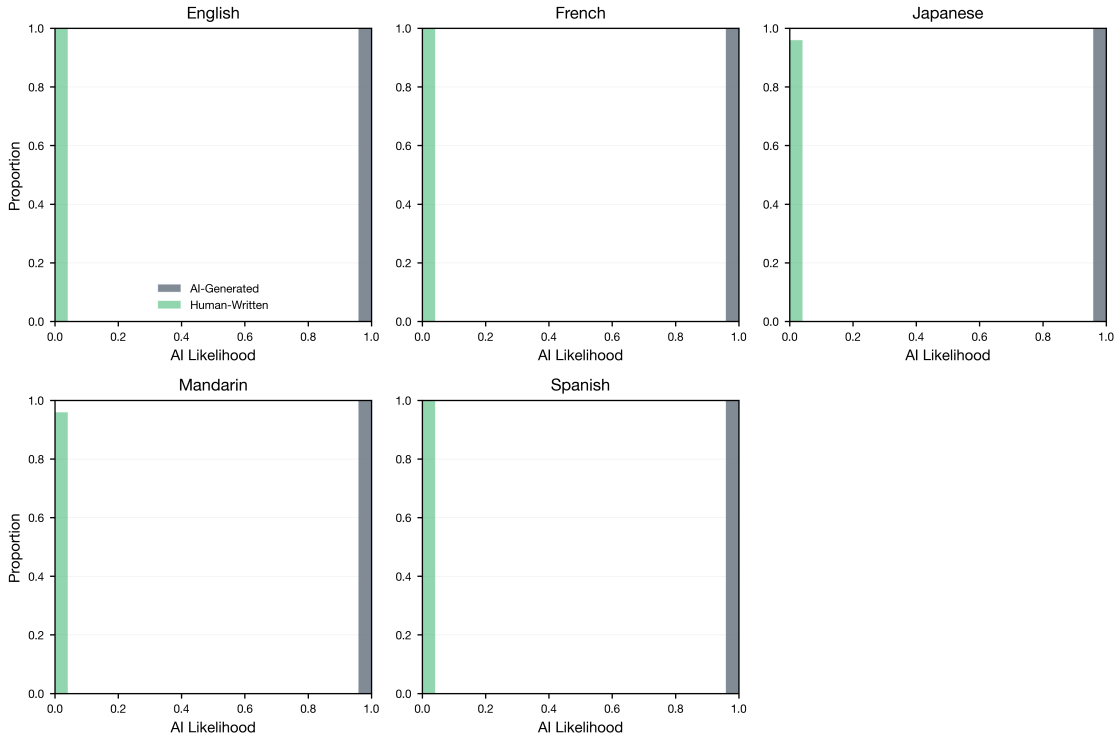


(b) Desklib



(c) DivEye

Figure 8 | **Language Robustness.** Comparison of detector score distributions across languages for Binoculars, Desklib, DivEye, and Pangram v3 (2/3).



(d) Pangram v3

Figure 8 | **Language Robustness.** Comparison of detector score distributions across languages for Binoculars, Desklib, DivEye, and Pangram v3 (3/3).

Summary. Based on this evaluation, we selected Pangram v3 as the primary detector for our analyses, since it came out as the most consistent option. By contrast, Binoculars and DivEye proved to be rather unreliable. Desklib performed similarly in the text length and model family robustness analyses; it even outperformed Pangram v3 in model version robustness; however, it underperformed relative to Pangram v3 on HTML vs. plain text and language robustness, which we deem more important for this analysis. An additional advantage of Pangram v3 over the three other detectors is that it operates in a three-way classification scheme (AI-generated, AI-assisted, human) that provides richer signal than binary detection.

B. Fact-Checking Annotation Interface

The fact-checking annotation study (Hypothesis 2) was conducted using a custom web application built with Flask and deployed on DigitalOcean App Platform with a PostgreSQL database backend. Annotators were recruited through Prolific (see Section 3.4 for demographics and representativeness information) and accessed the application via a direct URL with their Prolific participant ID passed as a query parameter, as shown in Figure 9.

Upon entering the application, annotators were presented with an instructions page describing the task, the four verdict categories, and guidelines for evidence evaluation, whose ending is shown in Figure 10. The instructions emphasized thoroughness (spending at least 1–2 minutes searching per claim), use of reliable sources, recording of evidence URLs, and objectivity. Annotators were told that "Not Enough Evidence" was a valid verdict when reliable information could not be found after a thorough search.

The annotation interface presented claims grouped by source article. For each claim, the annotator was required to: (1) select a verdict from the four-category scheme (Supported, Refuted, Not Enough Evidence, Conflicting Evidence), and (2) provide a confidence rating on a 1–5 scale. Optionally, annotators could record evidence URLs and free-text notes. The annotation view for a single claim in displayed in Figure 11. The interface tracked time spent per claim and per article. Upon completing all assigned articles, annotators received a completion code to submit on Prolific, as shown in Figure 12.

The four-category verdict scheme was adapted from the FEVER (Thorne et al., 2018) and AVeriTeC (Schlichtkrull et al., 2024) annotation frameworks. Claim assignments were managed by the application to ensure approximately 20% overlap across annotators, enabling computation of inter-annotator agreement via Krippendorff's alpha (Krippendorff, 2004), Fleiss' kappa, and Cohen's kappa for pairwise comparisons.

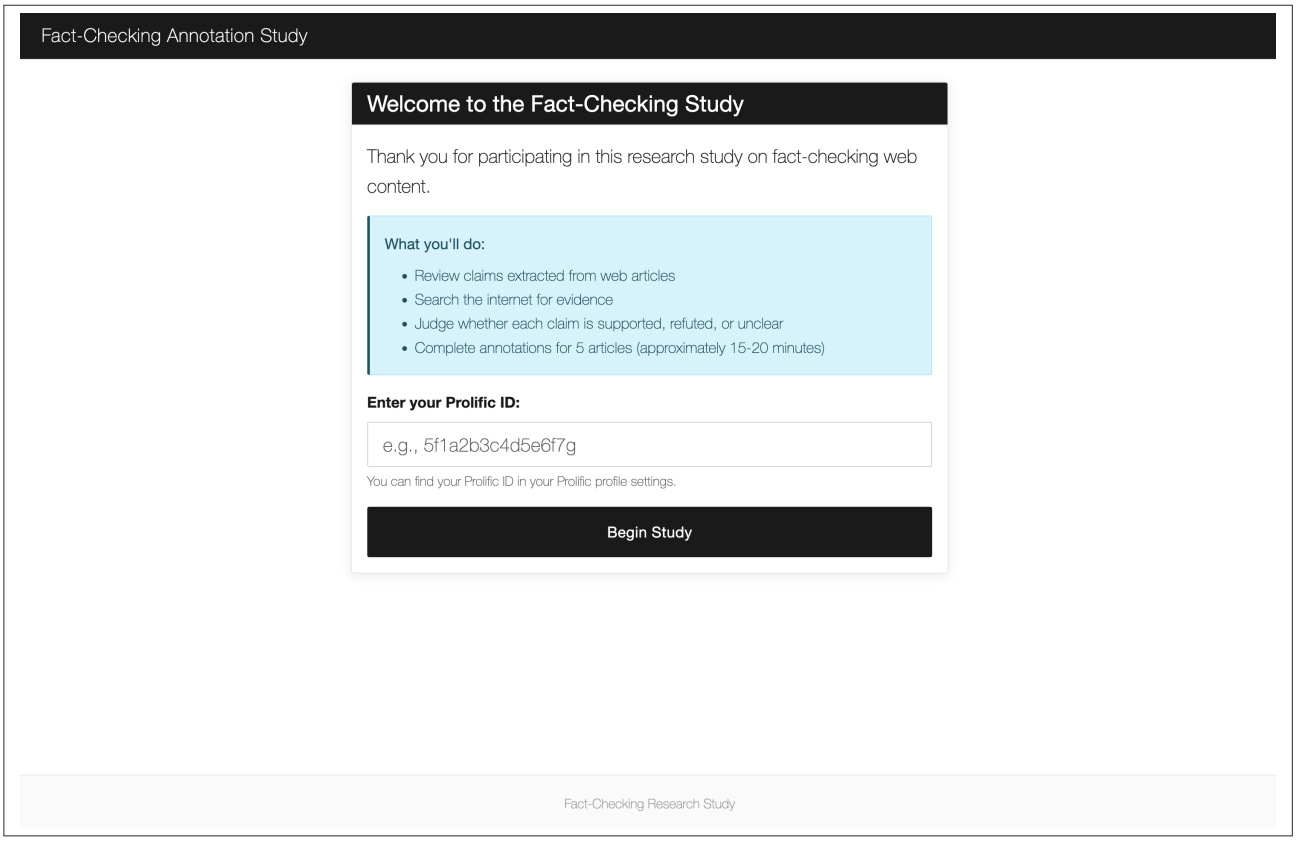


Figure 9 | **Welcome Screen.** The fact-checkers are presented with a list of expectations and asked to provide their Prolific IDs.

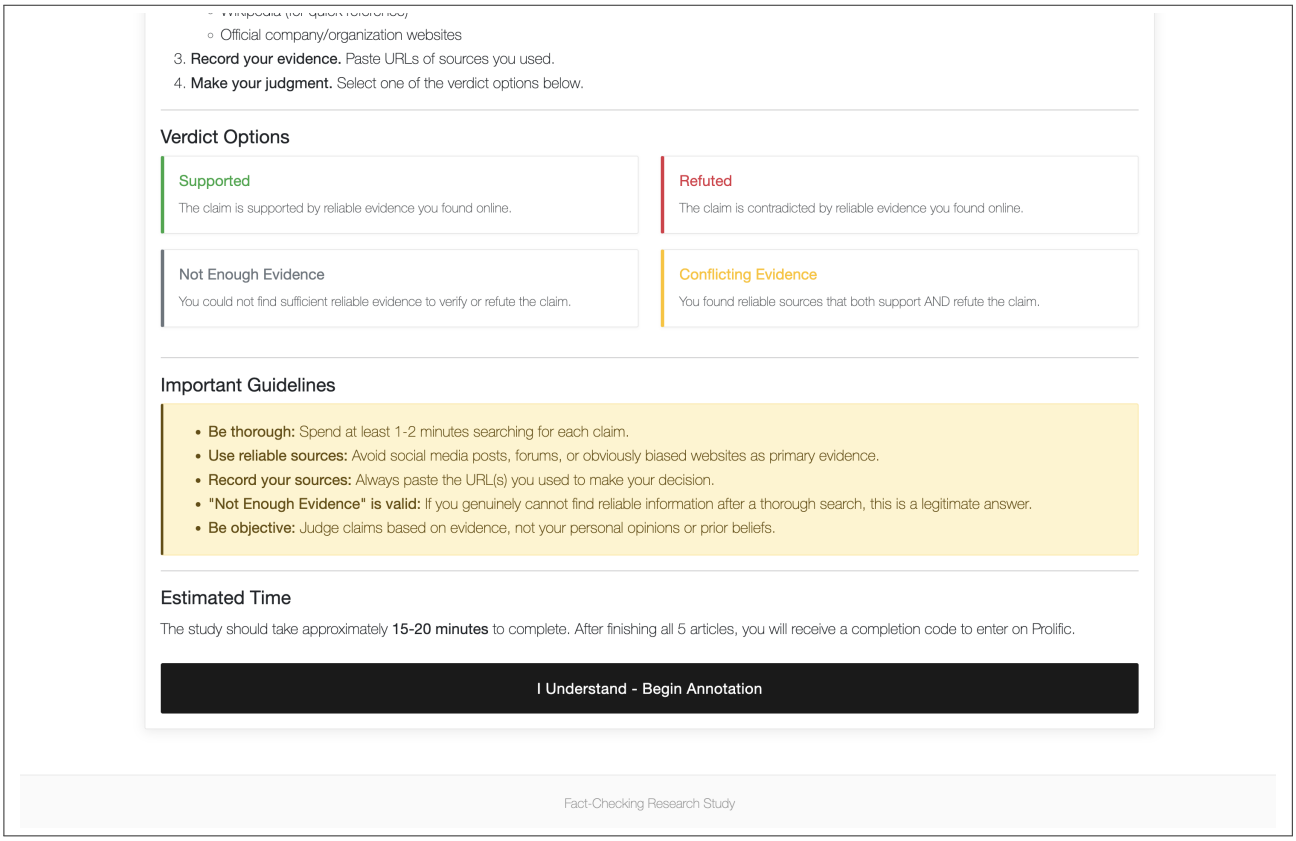


Figure 10 | **Instruction Recap Screens.** The fact-checkers are required to read through detailed instructions. The bottom of the screen is pictured.

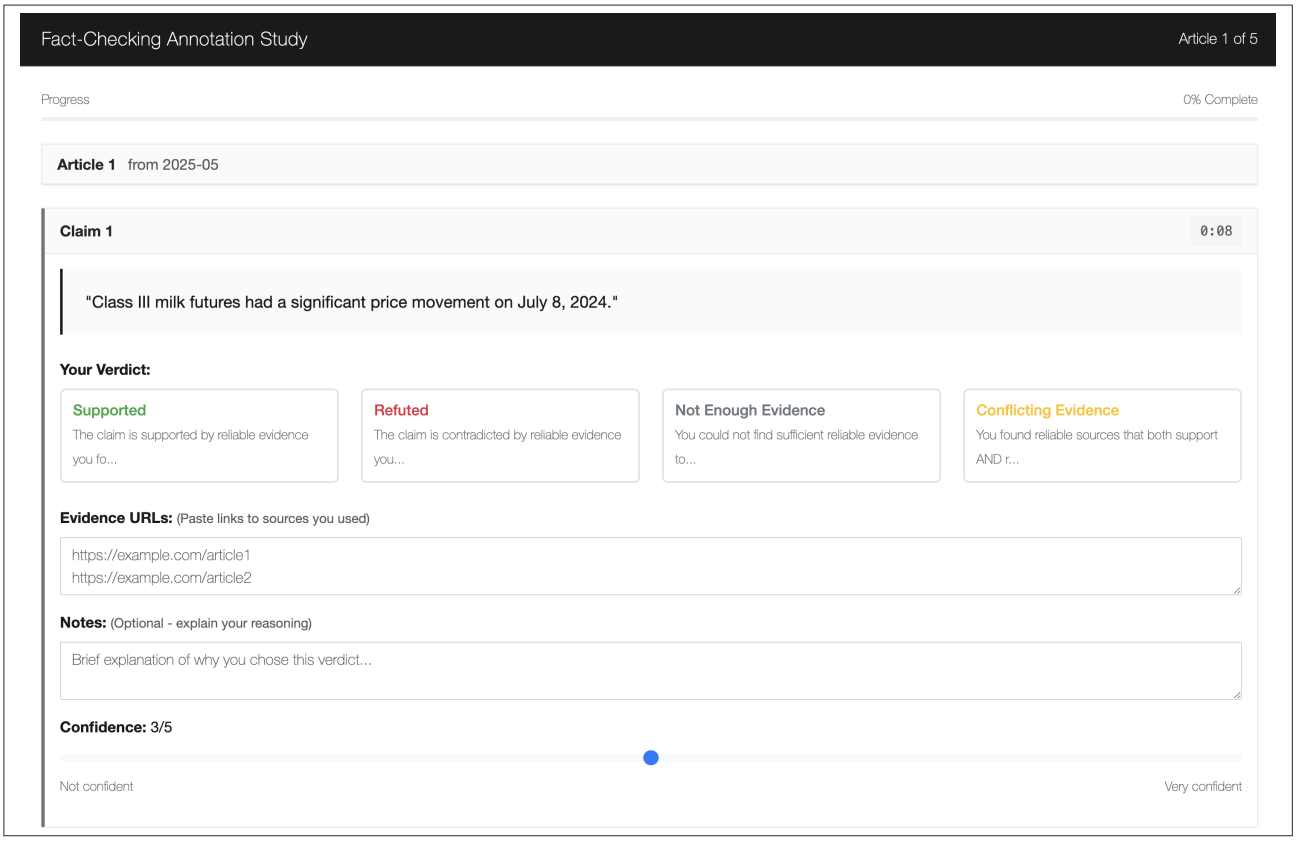


Figure 11 | **Item Annotation View.** Each claim for an article asks the fact-checkers to provide the verdict (Supported, Refuted, Not Enough Evidence, or Conflicting Evidence), evidence URLs, optional explanation, and confidence (range between 1, least confident, and 5, most confident). The top of the view indicates the progress within the assigned batch of articles.

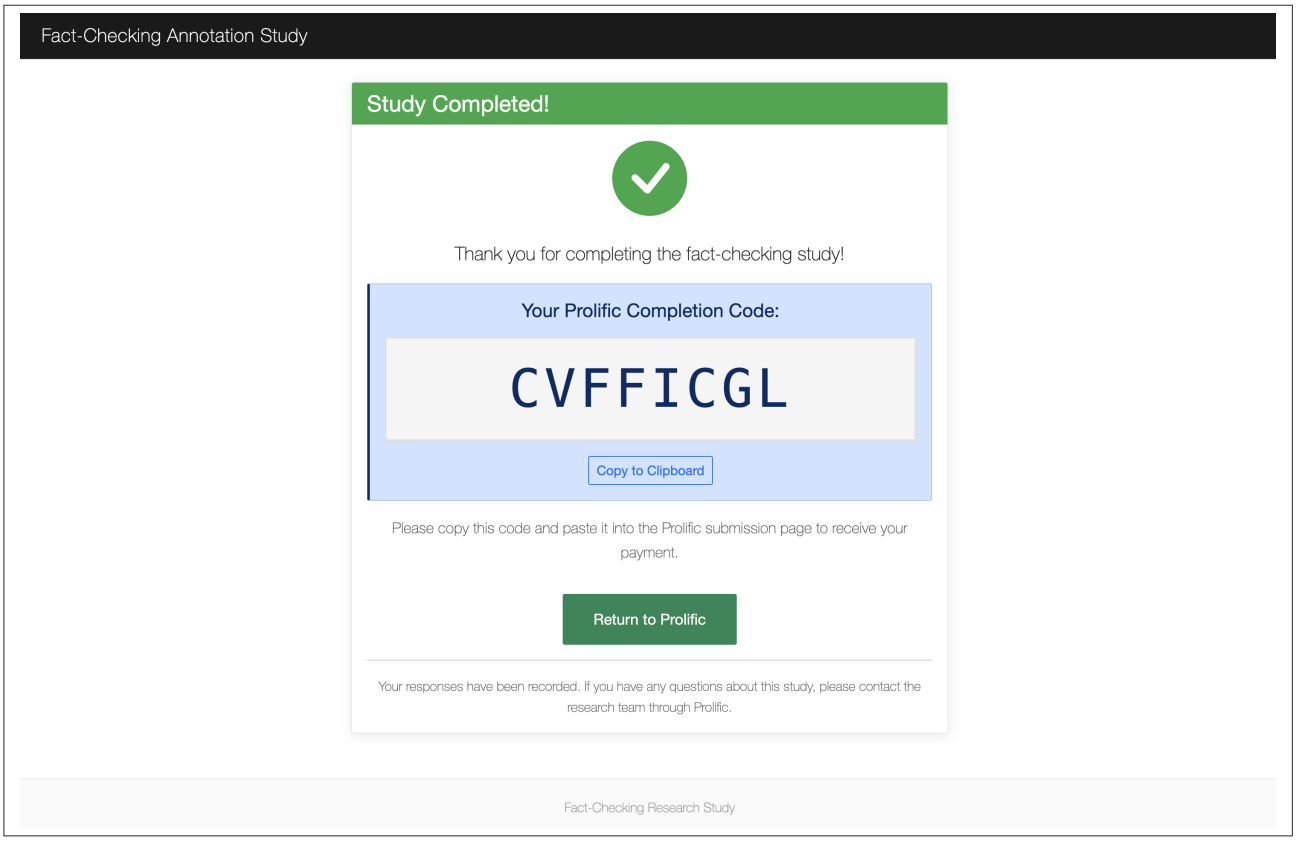


Figure 12 | **Completion Screen.** Upon completion, the fact-checkers are presented with the confirmation code for Prolific and instructions asking them to return to Prolific.

C. Additional Hypothesis Results

This appendix presents the full set of quantitative analysis and participant study figures for Hypotheses 2, 4, 5, and 6, which are summarized in the main text (Section 1) but whose figures are omitted from the main body for space. For each hypothesis, we show (a) the correlation between the measurable signal and the aggregate AI likelihood score across monthly samples, (b) the overall distribution of participant survey responses, (c) responses stratified by AI usage frequency, and (d) responses stratified by general view of AI's impact on society.

Hypothesis 2: Truth Decay

The Truth Decay Hypothesis posits that increasing AI-generated text on the internet leads to a higher rate of factually incorrect information. Figure 13 shows the quantitative analysis and participant study results. While 75.1% of respondents leaned towards agreement with this hypothesis, our quantitative analysis did not find a statistically significant correlation between the factual error rate and the aggregate AI likelihood score ($\rho = -0.19$, $p = 0.27$).

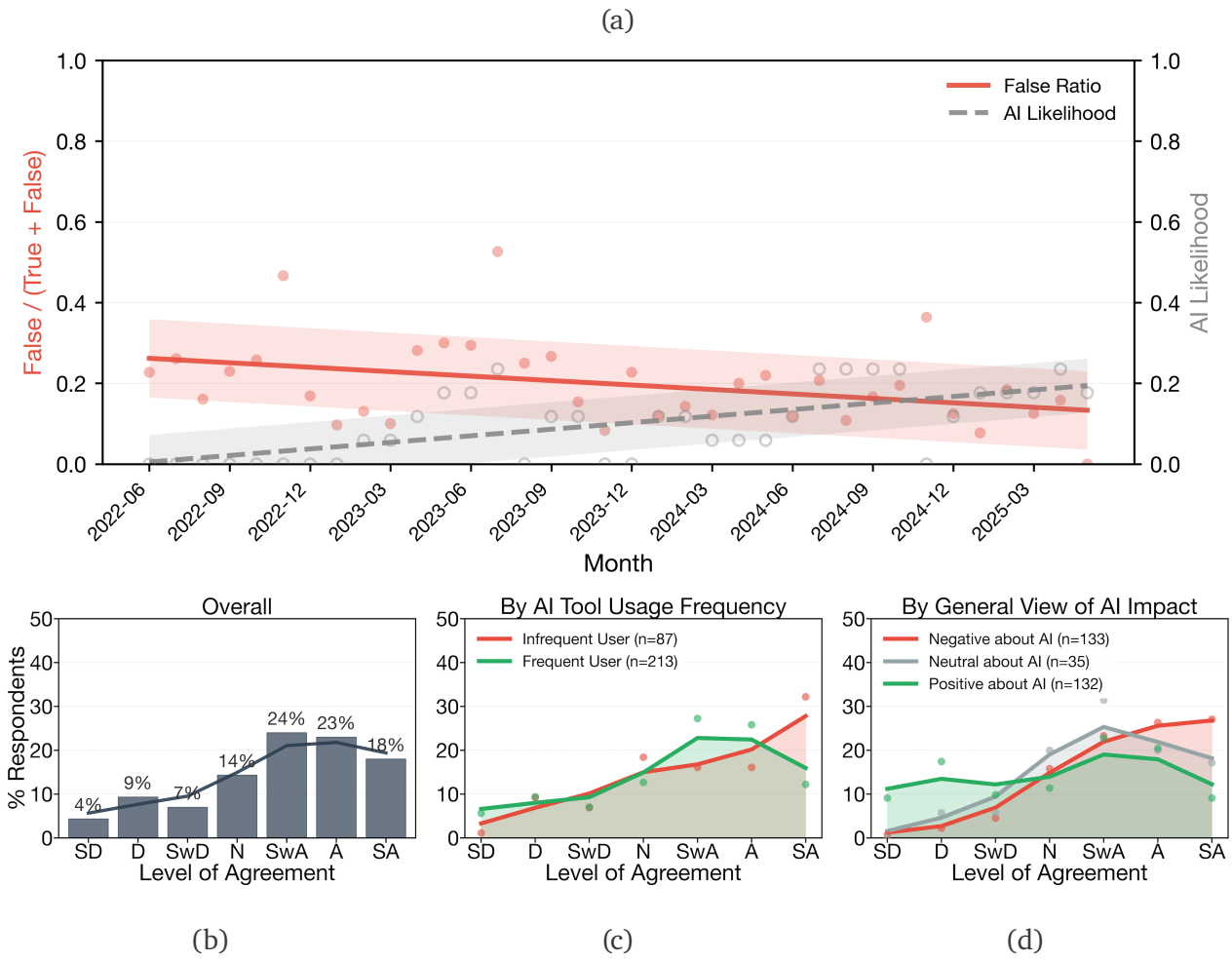
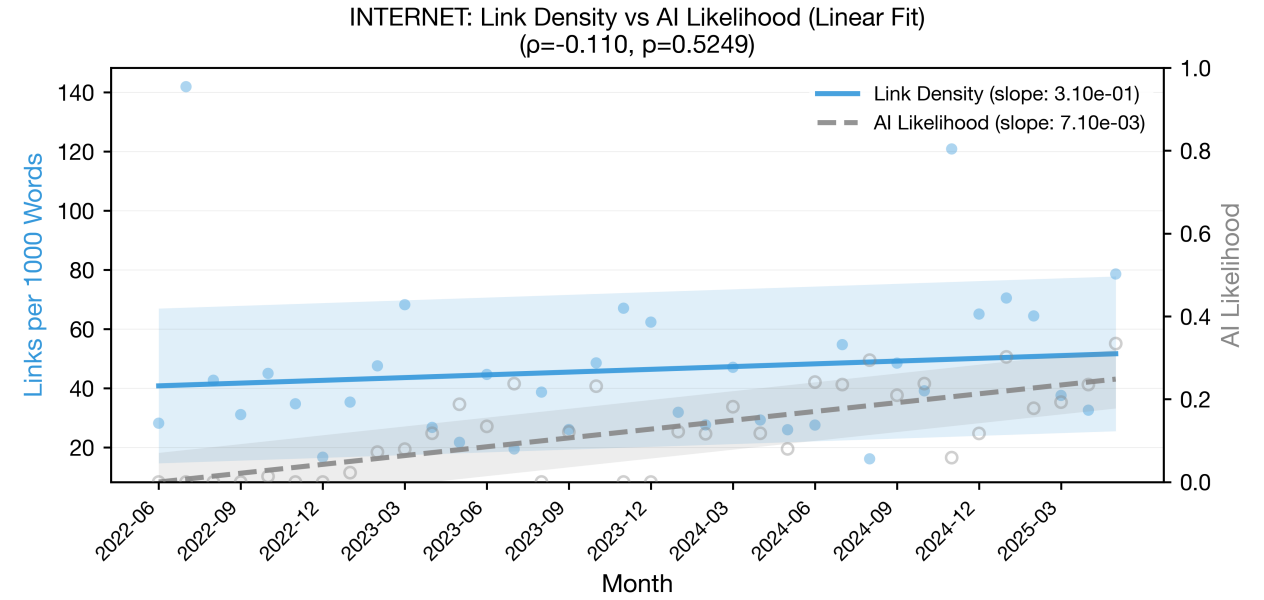


Figure 13 | **Results for Hyp. 2: Truth Decay.** The figure shows results for the Truth Decay Hypothesis from the participant study (RQ1) and quantitative analysis of randomly sampled websites from the Internet Archive (RQ3). In (a), the average factual error rate is plotted against AI Likelihood score, as detected by Pangram v3 ($\rho = -0.19$, $p = 0.27$). The overall results of the participant study are shown in (b), with responses ranging from Strongly Disagree (SD) to Strongly Agree (SA). These are broken down by AI usage frequency in (c) and general view of AI impact in (d).

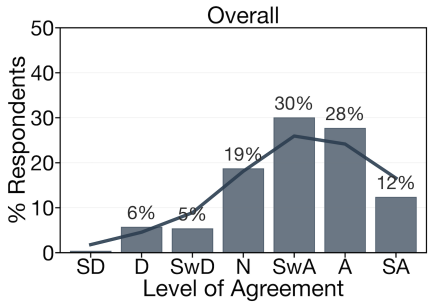
Hypothesis 4: Epistemic Islands

The Epistemic Island Hypothesis posits that as AI content becomes more common, articles increasingly provide answers without linking to external sources. Figure 14 shows the quantitative analysis and participant study results. While 69.9% of respondents leaned towards agreement, the quantitative analysis did not find a statistically significant inverse correlation between link density and the aggregate AI likelihood score ($\rho = -0.12$, $p = 0.48$).

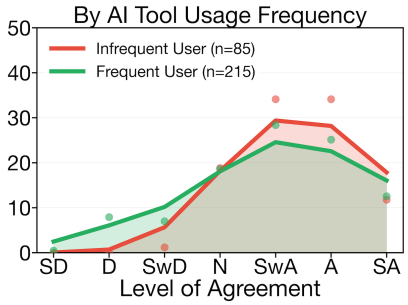
(a)



(b)



(c)



(d)

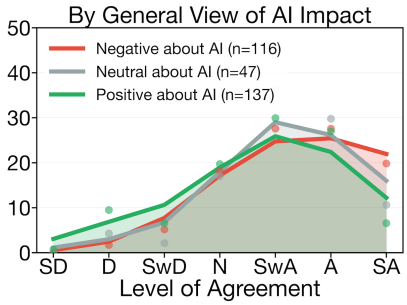


Figure 14 | **Results for Hyp. 4: Epistemic Islands.** The figure shows results for the Epistemic Island Hypothesis from the participant study (RQ1) and quantitative analysis of randomly sampled websites from the Internet Archive (RQ3). In (a), the outbound link density (links per 1,000 words) is plotted against AI Likelihood score, as detected by Pangram v3 ($\rho = -0.12$, $p = 0.48$). The overall results of the participant study are shown in (b), with responses ranging from Strongly Disagree (SD) to Strongly Agree (SA). These are broken down by AI usage frequency in (c) and general view of AI impact in (d).

Hypothesis 5: Entropy Dilution

The Entropy Dilution Hypothesis posits that as AI content becomes more common, content is becoming significantly longer while containing less actual meaning. Figure 15 shows the quantitative analysis and participant study results. While 60.7% of respondents leaned towards agreement, the quantitative analysis did not find a statistically significant correlation between the Gzip compression ratio and the aggregate AI likelihood score ($\rho = -0.02$, $p = 0.89$).

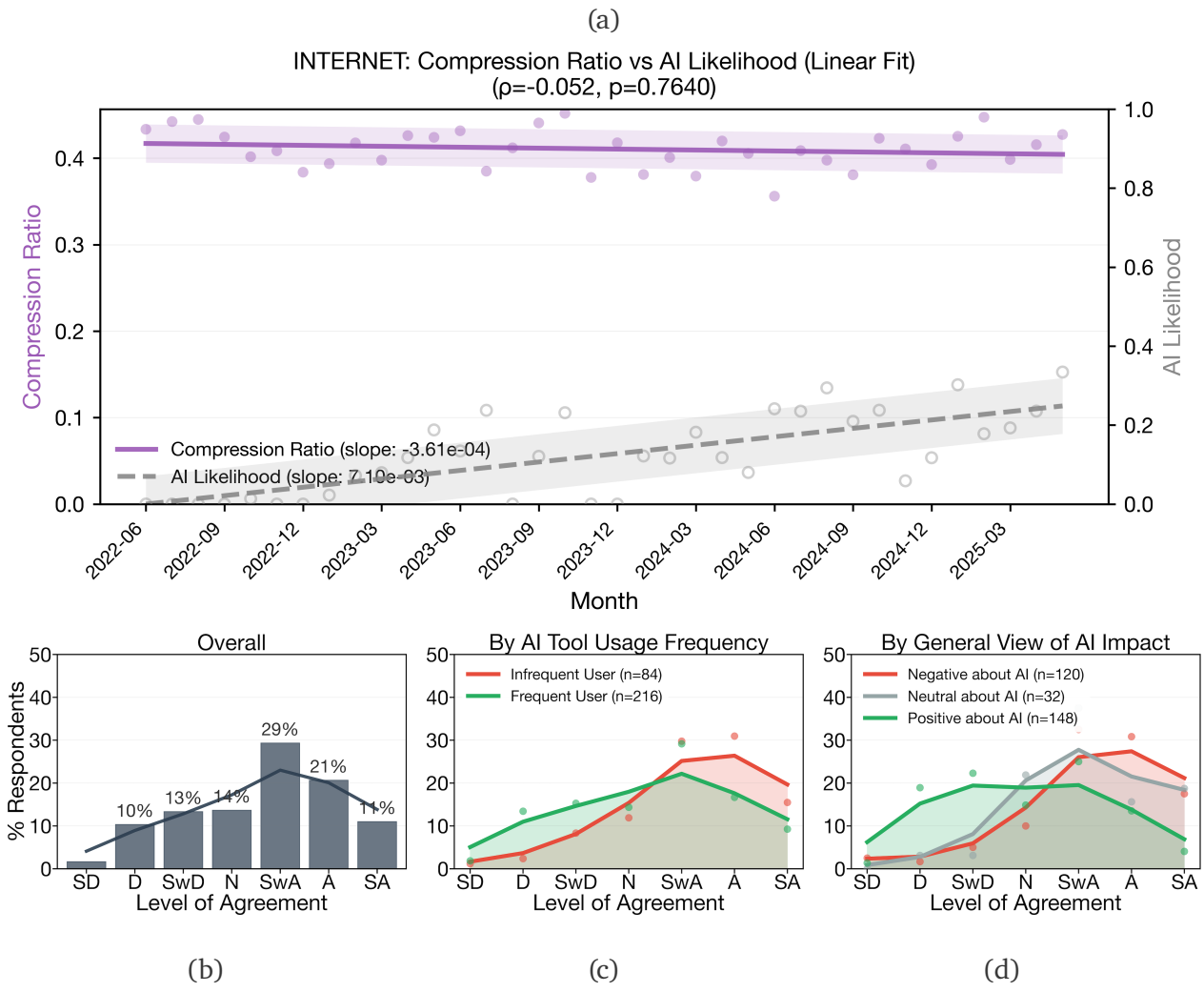
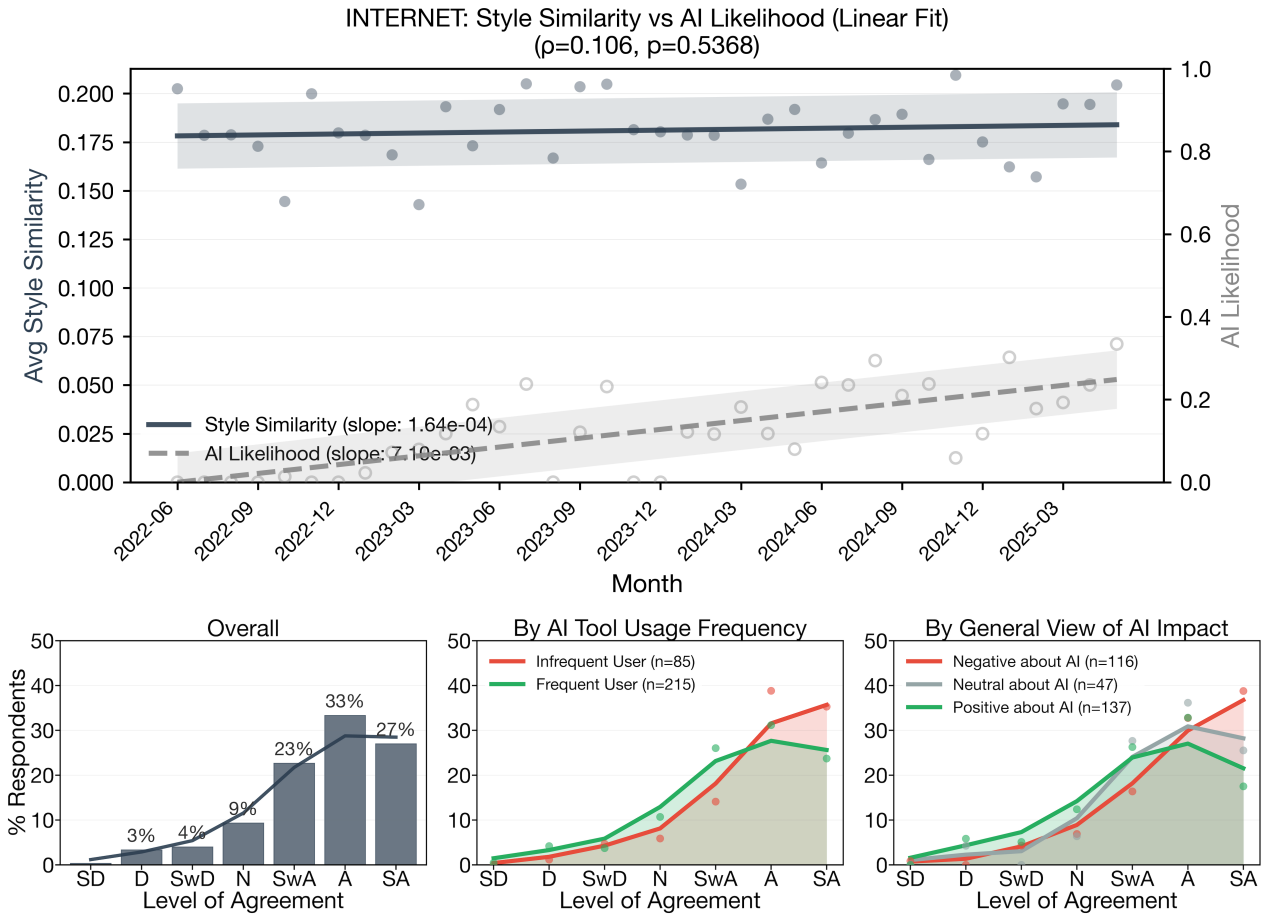


Figure 15 | **Results for Hyp. 5: Entropy Dilution.** The figure shows results for the Entropy Dilution Hypothesis from the participant study (RQ1) and quantitative analysis of randomly sampled websites from the Internet Archive (RQ3). In (a), the Gzip compression ratio is plotted against AI Likelihood score, as detected by Pangram v3 ($\rho = -0.02$, $p = 0.89$). The overall results of the participant study are shown in (b), with responses ranging from Strongly Disagree (SD) to Strongly Agree (SA). These are broken down by AI usage frequency in (c) and general view of AI impact in (d).

Hypothesis 6: Stylistic Monoculture

The Stylistic Monoculture Hypothesis posits that as AI content becomes more common, distinct individual writing styles are disappearing in favor of a generic, uniform voice. Figure 16 shows the quantitative analysis and participant study results. This hypothesis received the strongest agreement among participants (83.0% leaning towards agreement), yet the quantitative analysis did not find a statistically significant correlation between the average pairwise character 3-gram Jaccard similarity and the aggregate AI likelihood score ($\rho = 0.24$, $p = 0.17$).

(a)



(b) (c) (d)

Figure 16 | **Results for Hyp. 6: Stylistic Monoculture.** The figure shows results for the Stylistic Monoculture Hypothesis from the participant study (RQ1) and quantitative analysis of randomly sampled websites from the Internet Archive (RQ3). In (a), the average pairwise character 3-gram Jaccard similarity is plotted against AI Likelihood score, as detected by Pangram v3 ($\rho = 0.24$, $p = 0.17$). The overall results of the participant study are shown in (b), with responses ranging from Strongly Disagree (SD) to Strongly Agree (SA). These are broken down by AI usage frequency in (c) and general view of AI impact in (d).

D. Wayback Machine Sampling Details

Our sampling methodology follows Garg et al. (2025), with the following differences. First, we used a May 2025 ZipNum index and restricted our sample to URLs first archived on or after January 1, 2022, targeting the post-ChatGPT web rather than the complete web history. Second, we applied monthly temporal buckets with a target of 10,000 URLs per month, as opposed to yearly buckets of 1M URLs. Third, we applied a hard depth filter, excluding URLs with more than three path segments or more than two query parameters. Fourth, instead of logarithmic-scale downsampling, we applied a strict one-URL-per-host cap to prevent over-representation of frequently archived domains. Fifth, we explicitly required the first capture of each URL to return HTTP 200 with a `text/html` MIME type, verified via the CDX API. Finally, we shuffled each monthly bucket prior to sampling to neutralize the lexicographic ordering inherited from the ZipNum index. The last two steps were introduced because of the AI-generated text detection analysis.