DeGenTWeb: A First Look at MGT-dominant Websites

Sichang Steven He
University of Southern California
Los Angeles, USA

Calvin Ardi
Independent Researcher
USA

Ramesh Govindan
University of Southern California
Los Angeles, USA

Harsha V. Madhyastha
University of Southern California
Los Angeles, USA

arXiv:2605.00087v2 [cs.NI] 26 Oct 2026

Abstract

Many news reports have claimed that machine-generated text (MGT) from large language models (LLMs) is taking over the web. However, these claims are typically not based on a representative sample of the web, and the methodology underlying them is often opaque. Consequently, we lack an understanding of the true prevalence and characteristics of MGT on the web.

We describe DeGenTWeb, which systematically identifies MGT-dominant websites: sites whose informational prose content is mostly MGT, with little human input. We show how to adapt detectors of *MGT* for use on *web pages*, and how to aggregate detection results from multiple pages on a site for accurate *site-level* categorization. DeGenTWeb mislabels only 0.12% of human-authored sites as MGT-dominant. We find that 20.9% of the sites archived by Common Crawl between January and July 2026 and 13.6% of the sites linked from Bing's search results in 2025 are MGT-dominant. These figures are lower bounds on the prevalence of MGT-dominant sites. We also find that the fraction of MGT-dominant sites is growing over time, and that many such sites share the same owner. We will release all of our code and data after publication.

1 Introduction

Over three decades ago, Sir Tim Berners-Lee created the web to enable scientists to seamlessly share information with each other. Since then, the web has evolved to become arguably the largest distributed repository of human knowledge in history. Billions of websites provide access to a wide range of information, which we draw upon either directly via web searches or indirectly via the responses we receive from AI-powered chatbots.

However, today, the extent to which users can trust the information they discover on the web is increasingly under threat. This is because our assumption that the vast majority of web content has been authored by humans is being violated. In the current age of AI, it is natural for users to use LLMs to polish the prose they author. But, going beyond that, it is now also possible for a user to create an *entire website* with many pages simply based on a short textual prompt. For instance, given the title of a blog and a one-sentence summary of every blog post on it, state-of-the-art (SoTA) AI website builders [e.g., 1–5] can create all pages on the site.

In this paper, we focus specifically on sites whose pages primarily contain informational prose. Examples include sites which offer news articles, how-to guides and tutorials, descriptions of products and services, educational materials, technical documentation, etc. Problems with MGT, such as hallucination and plagiarism, are of particular concern on such sites because any user relies on their content to learn and make decisions. For example, malicious news and commentary sites have been known to use MGT to spread election-related disinformation [100]. We refer to sites where most, if not all, of the site's textual content is informational prose which has been generated using LLMs (i.e., is MGT) as *MGT-dominant*.

This observation that MGT-dominant websites are on the rise is not new. Several studies [9, 29, 37, 75, 81, 85] have attempted to quantify the prevalence of such sites. But all of these studies suffer from one of two fundamental drawbacks. Either the methodology used is opaque, e.g., many studies use proprietary MGT detectors, whose accuracy is unknown. Or the studies are conducted on a set of sites which are not representative of the web at large.

To fill this void, we present DeGenTWeb,[1] the first systematic methodology to identify MGT-dominant websites, and apply it to a large swath of the web. The primary challenge is that even the best MGT detectors are unreliable when configured to minimize the odds of falsely flagging human-authored text [31]. Moreover, classifiers of *chunks of text* cannot readily be applied to *web pages*, wherein multiple snippets of text are separated by HTML markup and boilerplate content is repeated across multiple pages on a site. In addressing these challenges, we make four contributions.

First, we show that the key to accurately identifying the use of MGT on the web is to categorize at the granularity of a *website* and to carefully select which subset of pages to analyze on each site. Though an MGT detector can be inaccurate on some pages, DeGenTWeb can accurately categorize a site by aggregating the detector's scores for multiple pages on the site. We perform this aggregation only across pages which, after we identify and exclude boilerplate content, contain a sufficient amount of prose for the detector to be reliable.

Second, to scientifically validate DeGenTWeb, we assemble a new ground truth dataset of 1,256 MGT-dominant and 10,328 non-MGT-dominant sites. We find that sampling 15 pages per site suffices to achieve accurate classification. We show that, via judicious filtering of pages and aggregation of the MGT detector's scores, we are able to achieve low false positive (0.12%) and false negative (∼2%) rates on our ground truth dataset.

Third, we use DeGenTWeb to classify 117k sites archived by Common Crawl and 18k sites found in Bing's results for 10k how-to search queries. We find that the fraction of MGT-dominant sites in Common Crawl's data has been steadily growing over the last few years: up from 2.0% in late 2022 to 20.9% in early 2026. We also find that MGT-dominant sites are highly prevalent in Bing's search results: for about two-thirds of our queries, the top 20 results point to at least one MGT-dominant site. These numbers are, however, lower bounds on the prevalence of MGT-dominant sites. This is

[1] Short for either Detect Generated Text on the Web, or Degenerate Web.

Table 1: The datasets we assemble for our study.

Dataset	#Sites	Source of data
CC2014	10,000	Common Crawl archive (Oct. 2014)
Body-swap	1,172	Page description → LLM → new page body
Polished	328	Page body → LLM → polished body
Wix/B12	45/39	Commercial AI website builders (2025)
New CC	117,069	Common Crawl archive (2020–Jul. 2026)
Search	18,169	How-to query → Bing → 20 results (2025)

because, in the interest of scalability and transparency, we use inexpensive open source MGT detectors, even though they fail to identify text generated by the newest LLMs [105, 108]. We find that DeGenTWeb would be able to identify these sites if paired with a proprietary detector, but that would cost over 300× more.

Lastly, we uncover the influence of financial incentives. Compared to other sites, MGT-dominant sites more often sell goods and services and are less often personal or organizational. We find that they are also less likely to attempt user retention (e.g., via subscriptions) and instead appear to monetize from one-off visits.

2 Datasets

Table 1 introduces the 6 datasets we use in this paper, totaling 147k sites[2] and 2.2M pages (~15 pages per site). We use the first four datasets to train and test DeGenTWeb, and the last two to quantify the prevalence of MGT-dominant sites in the wild.

2.1 Building an MGT-dominant baseline dataset

To train and test DeGenTWeb's website classifier, we need a ground-truth dataset where we know whether each site is MGT-dominant. There are many benchmark datasets for evaluating MGT detectors, but no such dataset exists for web pages or websites. To fill this gap, we compiled a dataset of 11,584 websites, totaling 174k pages in all, as follows.

Curating non-MGT-dominant sites ("CC2014"). We used HTMLs archived by Common Crawl (CC) [26], which is the largest publicly available sample of the web [13, 96]. Following existing common practices in curating human-authored text [31, 68, 98, 105], we sampled pages archived in 2014, before modern generative language models were prevalent in the public [38, 52]. To randomly sample sites archived by CC, we downloaded every 32nd of its October 2014 index files. From those, we selected records of type text/html with status code 2xx in which CC detects English text. For uniformly random yet reproducible sampling, we selected subdomains based on the BLAKE3 cryptographic hash [72] of every subdomain. For each selected subdomain, we uniformly sampled up to 640 pages, again based on BLAKE3 hash. Among the sampled sites with at least 15 qualifying pages each, we selected 10k; we describe later in §3.2 how we filter for pages with sufficient prose-like text.

Creating MGT-dominant sites ("Body-swap"). We lack reliable information as to which sites on the web are MGT-dominant. Therefore, we generated such sites ourselves. For every sampled page of each of 1,172 *CC2014* sites,[3] we swapped the main body of its HTML with MGT, leaving the rest untouched. We first used an LLM to briefly describe the original page. Then, in a fresh context with only that description, we used the LLM to generate the HTML with the new body. Each site used one of eight LLMs in turn: Claude Haiku 4.5, Sonnet 4, Sonnet 4.6, GPT-3.5 Turbo, GPT-4, GPT-OSS-120B, Llama 3.3 70B, and Mixtral 8x22B; we consider more recent models later in §4.3. Each *Body-swap* page retains the original page's metadata, layout, and boilerplate like sidebar and footer.

Creating non-MGT-dominant sites ("Polished"). We do not wish to classify a site as MGT-dominant if its pages have mixed human and LLM authorship, e.g., if a human writes the page's text first and then uses an LLM to polish it. Hence, just as we created the body-swap sites, we created 328 *Polished* sites. The only difference is that we directly gave LLMs the text from the original page and asked them to rewrite it (see §D for prompts).

Building out-of-distribution MGT-dominant sites ("Wix/B12"). To evaluate DeGenTWeb on MGT-dominant sites that are unlike our *Body-swap* dataset and better represent real-world websites, we used AI website builders from Wix.com and B12.io. Using sites of random Russell 2000 companies [11] as seeds, we asked ChatGPT to a) describe each website in one paragraph, b) suggest a name for a similar website, and c) write descriptions for 30 blog posts (see §D for the prompt used). We manually input each generated website's name and description into the web interface of Wix.com's AI website builder to generate the home page and boilerplate pages. We then input blog post descriptions to generate them one by one, randomizing each option: suggested blog title selection, blog length, style, and number of images. Similarly, we used B12.io to generate sites seeded from IndieWeb personal blogs [47] to produce a distribution distinct from the Wix sites.

In April 2025, we used this process to generate 45 Wix sites and 39 B12 sites. We also tried AI website builders Hostinger, Framer, 10Web, and Durable [2, 4–6]. But, their free trials were inadequate and their paid plans were beyond our means ($20 to $120 per site).

2.2 Sampling sites in the wild

To estimate the prevalence of MGT-dominant sites across the web, we studied sites from two sources: Common Crawl for a broad sample of the web, and Bing search results that capture what people searching for information see.

New CC Dataset. First, we randomly selected 1,297,346 of the sites archived by Common Crawl (CC) [26]. In the same way we curated the *CC2014* sites in §2.1, we uniformly sampled sites archived between January 2020 and July 2026, and then filtered out pages with insufficient non-boilerplate text (Table 2). We were left with 117,069 sites with at least 15 pages each.

Search Dataset. To study sites encountered by users, we issued 10k search queries to Bing between February and August 2025.[4] We crawled the sites which appeared in the search results. To mimic common how-to searches, each of our search queries used the title from a random wikiHow article [53]. The top 20 results for these queries collectively point to 59,046 sites. For each site, we sampled pages randomly either from the site's sitemap or from Wayback Machine's index for the site. For 18,169 sites, we were able to crawl 15 pages which passed our filtering criteria.

[2]We consider every fully qualified domain name as a unique website.

[3]The number of sites is constrained by our budget of $300 on OpenRouter.

[4]Unlike Google, results with Bing's search API matched its web search.

3 Detecting MGT-dominant Websites

To identify MGT-dominant websites, we leverage state-of-the-art detectors of MGT and address their limitations. We use Binoculars [41], a zero-shot detector which does not need to be rebuilt in order to detect text from every new LLM. Given a chunk of text, Binoculars outputs a score; the lower the score, the greater the likelihood the text was generated by an LLM. Binoculars has been shown to be accurate across many topical domains and source LLMs, is robust to many adversarial attacks, and is the most accurate at low false positive rates (FPR) [21, 31, 41, 55]; among the 15 MGT detectors we tried, we too find Binoculars consistently gives the lowest FPR (§4.2). We also find Binoculars to be accurate in differentiating MGT from LLM-assisted text (§4.1), consistent with prior studies [61].

3.1 Challenges and Approach

Identifying MGT on the web is, however, not as simple as executing Binoculars on every page. First, not all text on web pages is in the form of continuous prose, which MGT detectors expect as input; e.g., the text on a page may belong to a list of links or an interactive interface. Second, MGT detectors are less accurate when executed on short text, e.g., Binoculars' false negative rate (FNR) rises from 4% on text with 256 tokens to 20% on text with 128 tokens on one dataset [41]. Third, Binoculars offers a poor tradeoff between false positive and false negative rates; lowering FPR from 5% to 1% increases FNR from 20% to 27% on RAID [31] and from 33% to 42% in [97]. A low FPR is essential in the context of the web where, at least for now, human-authored prose dominates. For instance, if 5% of sites were MGT-dominant, a 1% FPR would mean that, even with no false negatives, roughly one-sixth of the sites classified as MGT-dominant would be incorrectly flagged as such.

We address these limitations based on two insights.

- **Aggregate page signals for robust site-level detection.** We observe that if the same source—LLM or human—produces many pages on a site, Binoculars' correct classifications eventually outnumber its misclassifications. Thus, for high accuracy, we opt to aggregate signals across multiple pages of the same site for a *site-level* classification. While some sites may have a mix of MGT and human-authored pages, our approach helps alert users about the quality of content on any site whose pages are consistently classified as MGT.
- **Focus on pages with enough non-boilerplate content.** We must carefully consider which subset of pages to analyze on any site and what portion of the content on each of these pages to consider. On the one hand, many websites include boilerplate content which is present on all of the site's pages. Hence, instead of feeding the text on any web page as is into Binoculars, we must identify and exclude such boilerplate text. On the other hand, we only consider pages which include sufficient content for Binoculars' classification to be reliable.

Below, we describe how we realize these insights in DeGenTWeb. Then, in §4, we validate our approach.

3.2 DeGenTWeb Detection Pipeline

As Figure 1 illustrates, to categorize a website, we must first crawl multiple pages from the site. For CC sites, we directly use their archived HTMLs. For others, we load each sampled page in a Chromium-based headless browser and save the page's final HTML. We then use Trafilatura [18], a SoTA web content extraction tool [19], to extract the main textual content from the HTML. It returns the content on a page that one would expect to see when the page is loaded in a browser's Reader mode, i.e., no boilerplate, recommendations, etc.

Page filtering. Because our focus is on informational prose and Binoculars is less accurate on shorter text [41], we only consider pages which have English content with 200 or more Falcon-7B tokens (used by Binoculars). Next, we discard pages which do not pass Dolma's quality filter [90]. Designed for LLM training data curation, Dolma's quality filter suits our exact needs to filter out non-prose text [90]. Following prior common practices [79, 80, 103], Dolma drops lines that do not end in punctuation, then rejects texts with heavy n-gram or line repetition, few alphabetic words, too few common English words, or too many lines that start with bullet points or end with ellipses, as detailed in §C. Lastly, we exclude pages on which more than half the page content is found on other pages on the same site, so that most of the text on each page we judge is unique to that site; we divide each page's content into chunks using Rabin fingerprinting [66] and detect exact duplicates of chunks of text.

Table 2 shows the pages our filters retain in each of the datasets that we collect either from CC or Bing's search queries. Pages are filtered out mainly because they have no English body text or are too short. In particular, because Common Crawl does not crawl pages with a headless browser, pages which are dynamically rendered have limited text in CC datasets. Moreover, we find that CC's classification of English pages is not always accurate.

Score aggregation. We execute Binoculars on the extracted text for every page that remains after filtering. For every site, we then compute the 10^{th}, 20^{th}, . . ., 90^{th} percentiles of the per-page Binoculars scores as a *feature vector* for the site. We input this 9-dimensional vector into a linear support vector machine (SVM; trained in §4) which outputs whether the site is MGT-dominant. We classify a site only when 15 pages remain after filtering; otherwise, DeGenTWeb returns no site label. We investigate our threshold choices in §4.

Table 2: Pages left (%) after each DeGenTWeb filter, and sites left, in the datasets from CC and Bing searches. Darker is more filtered out.

	CC2014	Search	New CC
All pages	2.54M	3.34M	50.31M
Has English text	62.2%	84.7%	92.1%
At least 200 tokens	23.2%	49.9%	32.0%
Passes Dolma filters	18.4%	44.3%	27.1%
At most 50% duplicated	15.3%	39.6%	20.7%
All sites	32.1K	44.1K	356.6K
Sites with 15 pages	10.0K	18.2K	117.1K

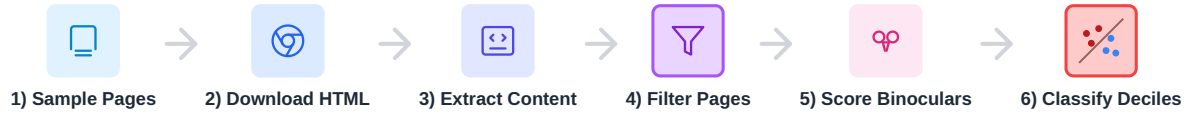


Figure 1: DeGenTWeb pipeline for detecting MGT-dominant websites.

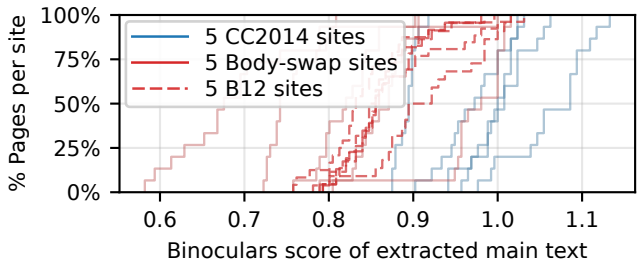


Figure 2: Distribution of per-page Binoculars scores for select ground-truth sites. Red is MGT-dominant.

4 Evaluating DeGenTWeb

We evaluate DeGenTWeb along several dimensions. First, we evaluate its accuracy on our datasets where we know the ground-truth, including in-distribution (ID) and out-of-distribution (OOD) sites. Second, we validate the need for and sufficiency of our approach via ablations, including replacing Binoculars with other MGT detectors. Lastly, we test DeGenTWeb on sites generated with newer LLMs and discuss accuracy versus cost tradeoffs.

We perform 30 trials in each experiment. In each trial, we randomly choose half of the *Body-swap* (586) and half of the *CC2014* (5,000) sites to train DeGenTWeb's SVM. We use each site's 9-decile page score feature vector (§3.2) and ground-truth label (MGT-dominant or not) to train the SVM. We then perform in-distribution (ID) testing on the remaining *Body-swap* and *CC2014* sites as well as sites in our *Polished* dataset, and out-of-distribution (OOD) testing on the sites we generate using Wix and B12. We see little difference in our ID results even when we attempt to maximize the difference between the training and test sets, i.e., for every *Body-swap* site used to train, the CC site from which it is derived is also in the training set. This is because our SVM has 10 trainable parameters, which cannot overfit on the hundreds of samples in the training data. We report the false positive rate (FPR)—what fraction of non-MGT-dominant sites we misidentify as MGT-dominant—and the false negative rate (FNR)—what fraction of MGT-dominant sites we miss. We focus on the 90th percentile (P90) of FPR and FNR on the test sets across all trials.

4.1 Accuracy with Binoculars

How accurate is DeGenTWeb? Across trials, DeGenTWeb achieves a P90 FPR of 0.12% on ID *CC2014* sites, and a P90 FNR of 1.5% on ID *Body-swap* sites. For the OOD sites generated using Wix and B12, DeGenTWeb still maintains a low P90 FNR: 2.2% on the 45 Wix sites and 2.6% on the B12 sites. Though the FNR is higher OOD than ID, it is unlikely that overfitting is the cause, for the reason mentioned earlier. Of the 328 Polished sites, DeGenTWeb only misclassifies 1 to 3 as MGT-dominant in each trial, resulting in a 0.61% P90 FPR.

Table 3: DeGenTWeb's site classification error rates as page filters cumulate.

Filtering condition	P90 FPR %	P90 FNR %
English	1.7	13.7
+ 200 tokens	0.82	6.0
+ Dolma	0.22	1.4
+ Deduplication	0.12	1.5

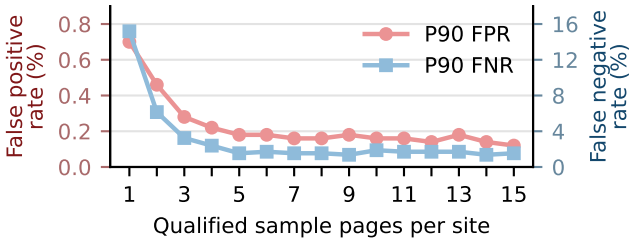


Figure 3: P90 FPR and FNR of DeGenTWeb when sampling varying numbers of pages per site for the SVM.

We inspected the 20 CC2014 sites that DeGenTWeb classifies as MGT-dominant. We find two dominant causes for these false positives. First, Binoculars, and MGT detectors in general, are prone to misclassifying common text or sentence patterns that they memorized during training [41] as the output of LLMs. For example, DeGenTWeb classifies htmlbible.com as MGT-dominant because that site hosts the Bible, which is often included in the data used to train LLMs.

Second, on sites like shopping1.hp.com, the body extracted by Trafilatura retained boilerplate text such as the disclaimer footers.

Is Binoculars by itself sufficient? Across select *CC2014*, *Body-swap*, and B12 sites, Figure 2 shows that the range of Binoculars scores overlaps across MGT-dominant and non-MGT-dominant sites. As a result, if we solely relied on classifying pages based on whether their Binoculars scores are below a threshold, we can correctly classify the text on only 87.0% of the pages on MGT-dominant sites as MGT. If we trained DeGenTWeb's SVM using a single ($N = 1$) page per site, we get a low P90 FPR of 0.70%, but a high P90 FNR of 15.2%. As we increase N in Figure 3, P90 FPR and FNR fall steeply until $N = 5$, then oscillate and stabilize at 0.12% and 1.5% when $N = 15$.

Is filtering of pages necessary? We study our choices of the various steps and associated thresholds we use in filtering pages. Table 3 shows that each of the measures we use for filtering pages

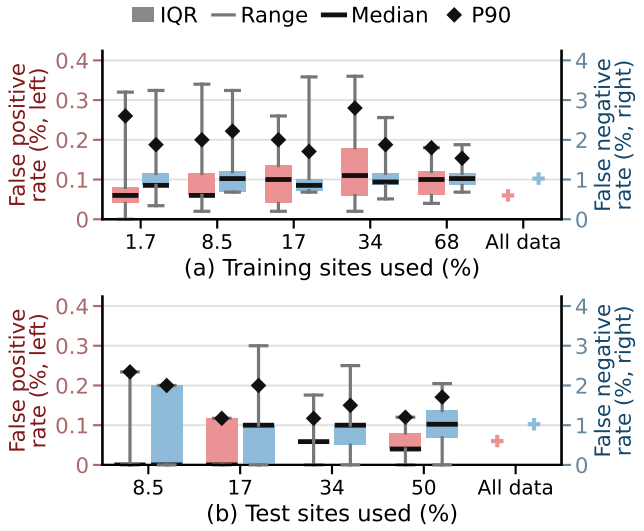


Figure 4: DeGenTWeb's error rates when varying the size of either (a) the training set or (b) the test set.

helps decrease FPR and FNR, from 1.7/13.7% when allowing all English pages to 0.12/1.5% after adding the final deduplication step. We experiment with different thresholds for the minimum fraction of non-duplicate content and the minimum number of tokens that a page must contain. We varied the former from 10% to 100% in steps of 10%, and the latter from 50 to 400 in steps of 50. None of these settings lowers P90 FPR below 0.12%; most increase it. Thus, we conclude that choosing different thresholds would not substantially change our results.

Is our ground-truth dataset large enough? We previously measured DeGenTWeb's accuracy by randomly splitting the *CC2014* and *Body-swap* sites in each trial into two equal partitions comprising the training and test set of sites. Now, we repeat the same experiment, wherein we continue to test on half of these sites, but use a random sample of the remaining half to train DeGenTWeb's SVM. We see little change in FPR and FNR as the training set grows (Figure 4a): P90 FPR is at most 0.28% at every size, and P90 FNR is 1.9% with 10 training *Body-swap* sites and 1.5% with 400. We also experiment with the full training set, and subsample the test set. Figure 4b shows that neither P90 FPR nor P90 FNR increases as the test set grows. Thus, we do not expect a significant change in DeGenTWeb's error rate with a larger training or test dataset.

4.2 Other MGT detectors

How does the choice of MGT detector affect accuracy? Our results thus far were based on DeGenTWeb's SVM using a vector of scores computed using Binoculars. We now experiment with several other MGT detectors: Fast-DetectGPT [16], L2D [111], FastNPR [86], DetectLLM-LRR [92], RAIDAR [67], BARTScore [112], Entropy [68], IRM [63], RoBERTa [89], and RADAR [44]. We also test Log Rank, Log Probability, and Rank [68], Max Probability [86]. Table 4 shows that all other options offer comparable or worse results. We also tried augmenting Binoculars with 4 other detectors, as well as all 15 detectors together. Each k-detector combination uses all 9 decile scores per detector to form a $9k$-dimensional feature vector. Since FPR and FNR depend on the threshold chosen, we show the standard AUROC for page-level accuracy. The only combinations that clearly beat Binoculars contain L2D, a rewrite-based detector which needs autoregression and costs ~6× as much as Binoculars in terms of GPU expenses in our tests.

Table 4: Page-level AUROC (%), and DeGenTWeb's site-level error rates (%) with different MGT detectors. Darker is worse.

Detector	Page AUROC	P90 FPR	P90 FNR
Fast-DetectGPT	98.4	0.18	1.4
Binoculars	98.3	0.12	1.5
L2D	97.8	0.22	2.7
FastNPR	92.7	0.80	13.0
DetectLLM-LRR	90.8	2.0	35.8
RAIDAR	88.6	0.88	17.1
BARTScore	88.3	2.0	31.2
Entropy	85.4	2.3	54.8
IRM	75.2	0.24	61.8
OpenAI RoBERTa	67.4	0.84	68.4
RADAR	67.0	1.9	72.9
Bino. + Fast-DetectGPT		0.30	1.5
Bino. + RAIDAR		0.18	1.4
Bino. + BARTScore		0.26	1.5
Bino. + L2D		0.040	0.68
All 15 combined		0.060	0.68

4.3 Newer LLMs and accuracy vs. cost tradeoff

We initially developed and validated DeGenTWeb in mid-2025. At that time, Binoculars enabled us to accurately classify the sites we created using Wix.

In September 2026, Wix overhauled its site builder. When we generated 67 new Wix sites, we found that DeGenTWeb is unable to classify any of them as MGT-dominant. The text on the median page across these new sites has a high, human-like Binoculars score of 0.97, compared to 0.86 for the old Wix sites. Replacing Binoculars with other MGT detectors, we find that L2D now performs the best, but still has a P90 FNR of 43% on the new Wix sites.

Since we lack visibility into the changes in Wix's back-end, we used 9 newer LLMs released in 2026 to generate 7 body-swap sites with each model. Sites created using GPT-6.1 Sol, GPT-6 Sol, GPT-5.5, Claude Opus 5.5, and Kimi K3 evade DeGenTWeb entirely, again due to high median Binoculars scores of 0.941 to 0.977. DeGenTWeb has a high P90 FNR of 86% for sites created with Claude Sonnet 5.5, while sites based on MiniMax M2.5, GLM-5, and DeepSeek V3.2 still yield lower P90 FNRs ranging from 0% to 29%. This is no surprise, as studies have shown that open-source detectors struggle to detect newer LLMs [e.g. 105, 108].

In contrast, in line with others' findings [48, 83], we find that Pangram's proprietary MGT detector [73] enables DeGenTWeb to identify all of our new Wix sites as MGT-dominant. Due to high cost, we sampled 660 Wix blog posts and 660 CC2014 pages. While

Table 5: Main findings about MGT-dominant sites.

§	Finding
5.1	Their share rose among newer websites
5.2	They are common in Bing's results for how-to queries
5.2	Some old sites are repurposed or switch to MGT
5.3	They typically have clear financial incentives
5.3	Monetization favors one-off visits over reader retention
5.3	Some of them share owners

using any single threshold on Binoculars' score would misclassify at least 37.5% of these pages, Pangram correctly classified all 1,320 of them.

Our small-scale experiment with Pangram already cost ~$100. So, using it for web-scale studies (§5) would be prohibitively expensive; classifying 100k sites would cost over $100k. In comparison, when running Binoculars with FP8-quantized Falcon-7B LLMs on an H100 GPU rented from modal.com in 2025, classifying 100k sites with DeGenTWeb cost about $300. Therefore, we accept the tradeoff of using Binoculars—low cost, but may miss newer LLM generations—and we treat our quantification of the prevalence of MGT-dominant sites in the next section as a conservative lower bound. If and when the use of Pangram at scale becomes economically viable, future work should explore using it as the building block for DeGenTWeb.

5 Findings in the Wild

Next, we use DeGenTWeb to study the prevalence and characteristics of MGT-dominant sites in our *New CC* and *Search* datasets (§2.2). These datasets are representative of the general open web and of the sites that users likely visit, respectively. Table 5 summarizes our main findings.

5.1 Rise of MGT in Common Crawl

How much of the New CC dataset is MGT-dominant? DeGenTWeb classifies 7.3% (8,526/117,069) of the *New CC* sites as MGT-dominant. We find that newer sites are much more likely to be MGT-dominant. Figure 5 shows that, since ChatGPT's launch, the share of MGT-dominant sites has risen steadily from 2.0% in the second half of 2022 to 25.0% in the first half of 2025. Since we cannot reliably date each site, we use the earliest archival date among our sampled pages from the site as a proxy for when the site was created. In the second half of 2025, our data show a dip. Given our results from §4.3, we speculate that this might be due to the adoption of newer LLMs in generating websites. Moreover, the number of sites in our dataset which were first archived in the first half of 2026 is limited (2,839 sites), which may affect the reliability of our result for that period.

Did sites transition to LLM content? Given our methodology (§3.2), DeGenTWeb will not classify a site as MGT-dominant if the text on some, but not an overwhelming majority, of the pages on the site appear to be MGT to Binoculars. This might occur if the pages on a site were previously being authored by humans, but are now generated using LLMs.

To evaluate the prevalence of this possibility, we analyzed the 34,172 sites in our *New CC* dataset for which we have at least 4

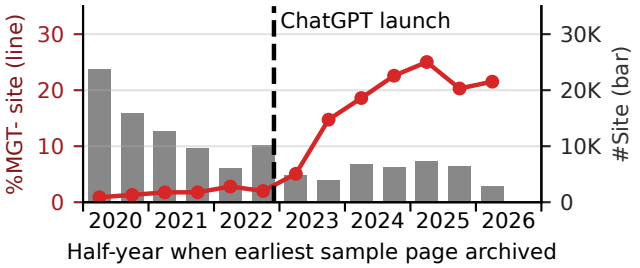


Figure 5: Share of MGT-dominant sites (line) and number of sites (bars) in the *New CC Dataset*, by half-year.

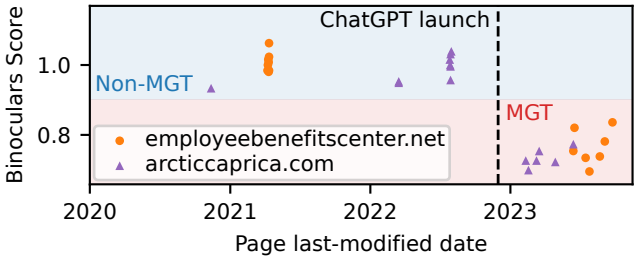


Figure 6: Examples of sites switching to MGT.

pages dated before ChatGPT's launch and at least 4 pages after; we use the Htmldate package [17] to extract every page's last-modified date. Since publishers may purposefully include incorrect dates and Htmldate is not perfect, we discard 2.5% of pages for which we clearly have the wrong date, i.e., those later than the date Common Crawl captured the page, and those Htmldate takes from a copyright notice. For 4,774 (14.0%) of these sites, the 75th percentile Binoculars score among the post-ChatGPT pages is less than the 25th percentile score among pre-ChatGPT pages. This large number is unlikely to be coincidental because, when we randomly shuffle the dates of each site's sampled pages 200 times, on average, we find that only 334 sites meet the same criteria.

Figure 6 shows two examples of such sites which appear to have transitioned from human-authored to LLM-generated text. For example, employeebenefitscenter.net used the same author name over the entire period and kept posting on the same topics, but its later posts are more formulaic and shorter (median 373 vs. 474 words before). In general, sites switching to MGT often also redesign: 20.1% of such WordPress sites changed their theme, vs. 14.0% among non-MGT-dominant sites. Few, however, changed their content management systems (3.4% vs. 1.3%).

5.2 MGT is Common in Search

How common are MGT-dominant sites in search results? For 43.8% (4,380/10,000) of our search queries, at least one of the top 10 results points to an MGT-dominant site; this fraction is 63.5% for the top 20. In total, DeGenTWeb classifies 13.6% (2,470/18,169) of all the sites in our *Search* dataset as MGT-dominant. This prevalence is 6.3 percentage points above that in the *New CC* dataset, likely because all of our Bing searches were in 2025 whereas our samples

from CC span from 2020 to 2026; sites within recent samples are more likely to be MGT-dominant.

Do MGT-dominant sites rank lower in search results? Bing does not seem to significantly down-rank MGT-dominant sites: their per-site median rank is 12.0 vs. 11.0 for non-MGT-dominant sites; average rank is 11.7 vs. 11.2. MGT-dominant sites appear about as often as non-MGT-dominant sites: an average number of links per site is 6.1 versus 6.0. This lack of discrimination may lower user experience if MGT-dominant content exhibits issues like hallucination and plagiarism.

What are MGT-dominant sites on old hostnames? Like in our *New CC* dataset, we find that many MGT-dominant sites which appear in Bing's search results existed pre-ChatGPT. Of the 2,470 MGT-dominant sites in our *Search* dataset, 1,669 (67.6%) have an archived page in the Internet Archive prior to 2023. We find, as discussed below, that these sites either are repurposed from previously inactive ones, or have switched to MGT and grown more quickly.

Repurposed. Of these sites, 19.5% (326) had fewer than 5 URLs archived by the Wayback Machine pre-ChatGPT and appeared unused or parked; 13.7% (229) have at least 5 before but had no archived pages in 2020–2022. These percentages are unlikely to be solely an artifact of the Internet Archive's crawling policy because the corresponding numbers for non-MGT-dominant sites are 1.6% and 4.2%. Examples of sites which have been repurposed to use MGT include HackySackGame.com (showed a parked domain page in 2013 but later hosted injury advice) and CGS.org (used to be the nonprofit Center for Governmental Studies in 1997, but later hosted generic blogs spanning topics from weather to tattoos).

Switched and grew. Of the remaining 1,114 (66.7%) sites, we consider the 883 sites with at least 3 pages archived both pre- and post-ChatGPT. If we use a threshold of 0.902 for Binoculars score (which maximizes its F1-score [41]), only 5.4% of the 7,998 pre-ChatGPT pages are classified as MGT, versus 65.3% of the 11,315 post-ChatGPT pages. For instance, post-ChatGPT, each of RAMIndustries.com's car-maintenance guides started with "`` `html ``", almost certainly from LLM-generated code fences.

Sites that switched to MGT likely published much faster, according to the differences in Wayback Machine capture rates. Among the 446 MGT-dominant and 1,389 non-MGT-dominant sites with at least 5 newly archived URLs every year in 2020–2022, MGT-dominant sites have a median of 2.3× as many newly captured URLs post-ChatGPT as pre-ChatGPT. The same ratio is 0.7× for non-MGT-dominant sites.

5.3 Characterizing MGT-dominant Sites

To better characterize each site beyond its usage of LLMs, we collect signals from 4 sources:

- Wappalyzer [77], which examines HTTP response headers and rendered HTML, to identify components of each site's **tech stack**, e.g., WordPress, Cloudflare, React
- EasyList CSS selectors [94] to identify **ads**
- **affiliate link** regex matching based on [20]
- an LLM (GPT-OSS-120B) to **categorize** each site based on example pages and metadata (see details in §E)

To minimize dollar costs, we extract signals from up to 3 pages per site. We do this for 10,070 MGT-dominant sites (whose pages

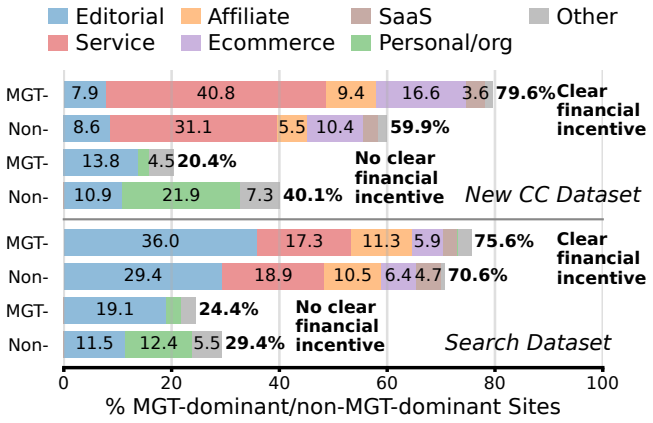


Figure 7: Category breakdown of MGT-dominant and Non-(MGT-dominant) sites, grouped by incentives.

we could fetch live, of all 10,996 detected) and 13,851 randomly sampled non-MGT-dominant sites. We use these signals to study these sites along two orthogonal dimensions as discussed below.

What incentivizes the creation of MGT-dominant sites? We partition the sites into two bins: those which have a clear financial incentive in attracting users to their site and those which do not. We perform this partitioning either based on the LLM's categorization of the site (e.g., offering a paid service, or serving affiliate/SEO content) or based on the presence/absence of ads and affiliate links on the site's pages.

Figure 7 shows that, in our *New CC* dataset, 79.6% of MGT-dominant sites have a **clear financial incentive**. The same fraction within non-MGT-dominant sites is 59.9%. In particular, two differences stand out. On the one hand, we see many businesses (service and SaaS sites) using LLMs to generate content for their websites. Many of these sites include a digital business card with an appointment form (e.g., HVAC and plumbing contractors, restoration services, and behavioral health clinics). Businesses catering to emerging markets (e.g., .ae, .pk, and .ph domains) are also common. On the other hand, personal/organizational sites are rarely MGT-dominant, likely the result of either not wanting to risk their reputation [22, 109] or lack of incentives.

Figure 7 also shows that, in contrast, the fraction of sites with a clear financial incentive in our *Search* dataset is similar in both MGT-dominant and non-MGT-dominant sites. This difference between the two datasets is because we only issue how-to queries, or because monetized sites rank equally well, with or without LLMs.

The one property that is shared with our *New CC* dataset is that personal/organizational sites are seldom MGT-dominant.

Interestingly, for a significant fraction of MGT-dominant publisher/editorial sites—13.8% in the *New CC* dataset and 19.1% in the *Search* dataset—we are unable to identify any clear financial incentives. Many of these sites seem to be low-traffic hobby projects, newly launched sites, or stale/defunct domains, e.g., niche single-author blogs on bird care, gemstones, and board-game rules, with no detectable monetization strategy (see §G).

How does monetization differ on MGT-dominant versus non-MGT-dominant sites? MGT-dominant sites lean towards entry-level ad stacks: AdSense (16.9% for MGT-dominant vs. 12.1% for non-MGT-dominant) and Ezoic (2.4% vs. 0.9%). Non-MGT-dominant sites more often use ad stacks which are open only to

high-quality sites or use mechanisms like header-bidding to maximize revenue

for publishers: Raptive, PubMatic, Sovrn, or Mediavine appear on 3.7% of MGT-dominant sites vs. 6.3% of non-MGT-dominant sites.

High-engagement monetization strategies are much less common among MGT-dominant sites. Subscription or donation links (e.g., Patreon, Ko-fi, Substack) are less common among them (3.1% vs. 8.6% on non-MGT-dominant), and audience-cultivation technologies (ConvertKit, Mailchimp, Substack , beehiiv, Ghost) appear on only 2.3% of MGT-dominant sites vs. 7.1% of non-MGT-dominant. Both gaps hold within every site category (§F). Together, these suggest that MGT-dominant sites tend to monetize one-off visits rather than a returning readership.

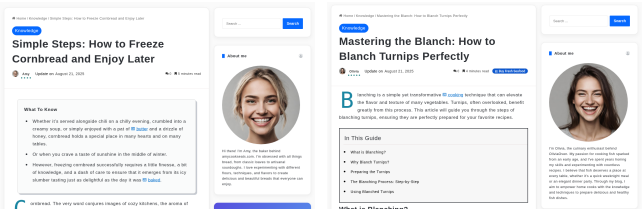
Are many MGT-dominant sites built by the same owner? We find strong evidence that many MGT-dominant sites have a shared owner. We identify shared owners by grouping sites which share AdSense publisher IDs or affiliate tags, or shared tech stacks with distinctive fingerprints.

Shared IDs or tags. The same AdSense ID means the same payee for ad revenue [36]; affiliate tags work similarly. Of 2,093 sites with AdSense IDs or affiliate tags, 256 (12.2%) use one of the 86 IDs or tags which appear on multiple sites; 171 of these sites are MGT-dominant. Sharing an AdSense ID is about twice as common among MGT-dominant sites: 17.3% vs. 8.8% for non-MGT-dominant sites. Interestingly, 24 of the 85 groups of sites linked by shared IDs and tags also include non-MGT-dominant sites. This suggests that some owners of multiple websites use LLMs to create some, but not necessarily all, of their sites.

The largest cluster comprises 18 MGT-dominant editorial sites under one AdSense ID. These sites span genres such as recipes, home, and lifestyle. Each site attributes every article to a single fictitious author with an AI-generated avatar and a generic bio. Figure 8a shows two examples. Each site in this cluster covers one narrow keyword, such as eggs, and has 8,358 to 26,316 posts (over 290,000 in all), likely to maximize ad views.

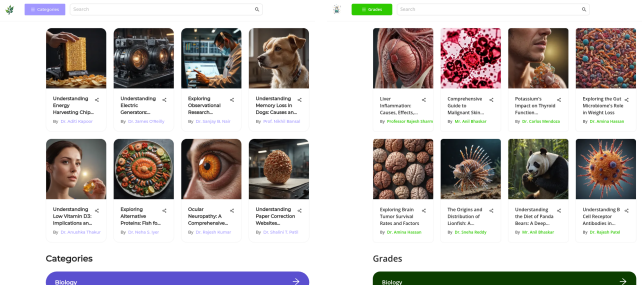
Identical tech stacks. When we cluster sites with identical tech stacks as detected by Wappalyzer [77], we identify 270 multi-site groups comprising 1,094 (10.9%) of the MGT-dominant sites. Tech stack alone does not indicate MGT-dominance: 93 groups (543 sites) share a stack with non-MGT-dominant sites. However, grouping sites with the same tech stack helps us identify the common underlying motivation, which may otherwise not be apparent from examining each site in isolation.

As an example, consider a group of 32 identical-looking MGT-dominant editorial sites with a unique fingerprint. All of these sites run Gatsby and serve an identical set of 3 unique GraphQL query hashes, which confirms shared source code. None of the 32 sites contain ads or affiliate links, but they all link to sites in the Russian network SAPE, which the linked-to sites pay for to boost their search rankings [34, 59]. We see that each site shows generic informational blogs focusing on one topic, e.g., lifestyle, jewelry,



amycookseats.com oliviaoven.com

(a) Two of the 18 sites that share an AdSense ID.



sciquants.com biocascades.com

(b) Two of the 32 sites with an identical tech stack.

Figure 8: Examples from two groups of MGT-dominant sites.

or science (tranquilyx.com, stongle.com, sciquants.com). Figure 8b shows two examples.

6 Related Work

MGT detectors. Recent surveys cover this huge body of work [8, 57, 104, 106]. For our purpose, statistical zero-shot detectors [e.g. 15, 16, 25, 41, 113] are considered to generalize better than learning-based ones [e.g. 14, 44, 50]; detectors based on text generation are more accurate but too expensive [42, 46, 69]. Manual validation is infeasible because humans are not reliable at detecting MGT [22, 35]; even SoTA detectors heavily miss MGT under low FPR [97] and produce many false positives on social web text [32]. Finally, we do not aim to address attacks for evading detection, many of which are effective [33, 101, 105]. Our methods build upon these text detectors and balance accuracy and practicality.

Web spam and search engine optimization (SEO). Web spam has long been characterized [39], detected [12, 23, 70, 91], and shown to compete in search ranking [28, 30] and generate ad revenue [102]. Spam sites commonly deploy SEO [60, 65, 87] and generative engine optimization (GEO) [7, 76], which retrieval engines constantly battle against [20]. MGT is found to outrank human text in question-answering [24] and web retrieval [107], and is found to be prioritized by neural retrievers [27] but not term-based retrievers [45]. While MGT-dominant sites may involve spam or SEO, we focus instead on their prevalence and characteristics.

MGT measurement in the wild. Broadly, studies have found MGT rising in research publications [54, 56, 62, 74], peer reviews [84, 110], crowd work [99], and malicious emails [43]. Specialized studies have found MGT in various corners of the web, e.g., its share is growing in fact-checked disinformation [64]. For online forums, studies found that MGT responses are prevalent in MOOCs [71], grew

sharply on Medium and Quora [93] but slowly on Reddit [58], and are reacted to much like human answers on a Q&A platform [88]. For news, studies found that MGT rose on mainstream news and misinformation sites [40], appears frequently in American newspapers [82], and is easy to produce and hard to detect [78]. Much of the web in lower-resource languages is also machine translated [95]. Our work complements these more specific studies and focuses broadly on the open web and user-facing web content.

Similar to ours, unpublished work surveyed the web to quantify MGT pages from Common Crawl [75, 81], AI-generated or AI-assisted sites in the Internet Archive [29], and MGT sources that generative engines cited [10]. However, they either use proprietary detectors [10, 29, 81] or a restricted-access detector with a much higher FPR of ∼1% [75]. In comparison, our methodology is scalable and fully transparent, while achieving low FPR by targeting site-level categorization of MGT-dominance, not identification of LLM-assistance or classification of individual pages.

7 Conclusion

It is widely presumed that the number of websites generated using LLMs is on the rise, but a transparent, reliable methodology to quantify such sites has been lacking. In this paper, we presented DeGenTWeb to fill this void. Specifically, we demonstrated how to employ state-of-the-art MGT detectors to reliably classify websites which have significant non-boilerplate text. We also used DeGenTWeb to study a broad cross-section of the web. As LLMs evolve, we believe that our methodology will continue to be applicable when paired with new detectors.

Acknowledgments

We thank the anonymous reviewers, Zubair Shafiq, Enze "Alex" Liu, Marco Mellia, and Jonathan David Louis May for their valuable feedback. Steven thanks Bhaskar Mahajan for their code contribution, Janek Bevendorff for their help, and Arvin Ghavidel, Yibo Yan, Yizhuo "Coulson" Liang, Avi Schwarzschild, Huijuan Wang, Taiwei Shi, and Zihao Feng for their feedback; special thanks to Shaoyu Wang and Huanchen Sun. Harsha is supported by NSF award 2403432 and an Amazon Research Award (Spring 2025). Steven is supported by Midas Intelligence Inc.

References

[1] 2026. AI Website Builder - Create a Website in Minutes | Wix. https://www.wix.com/ai-website-builder.
[2] 2026. AI Website Builder | Launch your dream site in seconds | Hostinger. https://www.hostinger.com/ai-website-builder.
[3] 2026. B12 | The easiest AI website builder. https://www.b12.io/.
[4] 2026. Durable – AI Business Builder | Launch in minutes. https://durable.com/.
[5] 2026. Framer: Free website builder and AI design agent. https://www.framer.com/.
[6] 2026. Launch and Grow Your Business Online with 10Web. https://10web.io/.
[7] Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
[8] Zainab Ahmad, Miguel Torres-Ruiz, Ahmad Mahmood, Rolando Quintero, Iqra Ameer, and Necva Bölücü. 2026. Human or Machine? A Survey on Machine-Generated Text Detection. *IEEE Access* (2026).
[9] Ahrefs. 2025. 74% of New Webpages Include AI Content (Study of 900k Pages). https://ahrefs.com/blog/what-percentage-of-new-content-is-ai-generated/
[10] Mowafak Allaham and Nicholas Diakopoulos. 2026. Synthetic Sources?: Auditing Generative Search Engine Citations for Evidence of AI-Generated Sources. https://arxiv.org/abs/2605.23684
[11] AnnualReports.com. 2025. Russell 2000 Index Companies. https://www.annualreports.com/FeaturedProgram/15
[12] Calvin Ardi and John Heidemann. 2019. Precise Detection of Content Reuse in the Web. *SIGCOMM Comput. Commun. Rev.* (2019).
[13] Stefan Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*.
[14] Sankalp Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. Fine-tuning Language Models for AI vs Human Generated Text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.
[15] Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. 2025. Glimpse: Enabling White-Box Methods to Use Proprietary Models for Zero-Shot LLM-Generated Text Detection. In *The Thirteenth International Conference on Learning Representations*.
[16] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *The Twelfth International Conference on Learning Representations*.
[17] Adrien Barbaresi. 2020. htmldate: A Python package to extract publication dates from web pages. *Journal of Open Source Software* (2020).
[18] Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*.
[19] Janek Bevendorff, Sanket Gupta, Johannes Kiesel, and Benno Stein. 2023. An Empirical Comparison of Web Content Extraction Algorithms. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
[20] Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. 2024. Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines. In *European Conference on Information Retrieval*.
[21] Janek Bevendorff, Matti Wiegmann, Emmelie Richter, Martin Potthast, and Benno Stein. 2025. The Two Paradigms of LLM Detection: Authorship Attribution vs Authorship Verification. In *Findings of the Association for Computational Linguistics: ACL 2025*.
[22] Bynder. 2024. Study reveals how consumers interact with AI-generated content vs human-made. *Bynder Press and Media* (2024).
[23] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. 2007. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*.
[24] Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024. Spiral of Silence: How is Large Language Model Killing Information Retrieval?—A Case Study on Open Domain Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
[25] Zhihui Chen, Kai He, Yucheng Huang, Yunxiao Zhu, and Mengling Feng. 2025. DivScore: Zero-Shot Detection of LLM-Generated Text in Specialized Domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
[26] Common Crawl. 2025. Common Crawl. https://commoncrawl.org/
[27] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural Retrievers are Biased Towards LLM-Generated Content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
[28] Cristian Danescu-Niculescu-Mizil, Andrei Z Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Bo Pang. 2010. Competing for users' attention: on the interplay between organic and sponsored search results. In *Proceedings of the 19th international conference on World wide web*.
[29] Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. The Impact of AI-Generated Text on the Internet. (2026).
[30] Ye Du, Yaoyun Shi, and Xin Zhao. 2007. Using spam farm to boost PageRank. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*.
[31] Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. 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*.
[32] Arka Dutta, Utkarshani Jaimini, Utkarsh Bhatt, Sara Shree Muthuselvam, Amitava Das, and Ashiqur Rahman KhudaBukhsh. 2026. A Large Scale Social Web Audit of AI Generated Text Detection Systems. *Proceedings of the International AAAI Conference on Web and Social Media* (2026).
[33] Hao Fang, Jiawei Kong, Tianqu Zhuang, Yixiang Qiu, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. 2025. Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors. In *Proceedings of the 2025 Conference on Empirical Methods in*

Natural Language Processing.
[34] Charles Floate. 2015. A Beginners Guide to Buying SAPE Links. https://www.charlesfloate.com/sape-links Accessed 2026-04-28.
[35] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. 2024. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)*.
[36] Google. 2026. Find your publisher ID. Google AdSense Help. https://support.google.com/adsense/answer/105516?hl=en Accessed September 14, 2026.
[37] Graphite.io. 2025. More Articles Are Now Created by AI Than Humans. https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans
[38] Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2017. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* (2017).
[39] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web Spam Taxonomy. In *AIRWeb*.
[40] Hans W. A. Hanley and Zakir Durumeric. 2024. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *Proceedings of the International AAAI Conference on Web and Social Media* (2024).
[41] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. In *Proceedings of the 41st International Conference on Machine Learning*.
[42] Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. 2025. Learning to rewrite: Generalized llm-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.
[43] Wei Hao, Van Tran, Vincent Rideout, Zixi Wang, AnMei Dasbach-Prisk, MH Afifi, Junfeng Yang, Ethan Katz-Bassett, Grant Ho, and Asaf Cidon. 2025. Do spammers dream of electric sheep? characterizing the prevalence of llm-generated malicious emails. In *Proceedings of the 2025 ACM Internet Measurement Conference*.
[44] Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-Text Detection via Adversarial Learning. In *Advances in Neural Information Processing Systems*.
[45] Wei Huang, Keping Bi, Yinqiong Cai, Wei Chen, Jiafeng Guo, and Xueqi Cheng. 2026. How Do LLM-Generated Texts Impact Term-Based Retrieval Models?. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*.
[46] Yifei Huang, Jiuxin Cao, Hanyu Luo, Xin Guan, and Bo Liu. 2025. Magret: Machine-generated text detection with rewritten texts. In *Proceedings of the 31st International Conference on Computational Linguistics*.
[47] IndieWeb community. 2025. IndieWeb. https://indieweb.org/
[48] Brian Jabarian and Alex Imas. 2025. *Artificial Writing and Automated Detection*. Working Paper. National Bureau of Economic Research.
[49] Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. 2022. The Internet with Privacy Policies: Measuring The Web Upon Consent. *ACM Trans. Web* (2022).
[50] Kaijie Jiao, Quan Wang, Licheng Zhang, Zikang Guo, and Zhendong Mao. 2025. M-RangeDetector: Enhancing Generalization in Machine-Generated Text Detection through Multi-Range Attention Masks. In *Findings of the Association for Computational Linguistics: ACL 2025*.
[51] Andrew J. Kaizer and Minaxi Gupta. 2016. Characterizing Website Behaviors Across Logged-in and Not-logged-in Users. In *Proceedings of the 2016 Internet Measurement Conference*.
[52] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and Understanding Recurrent Networks. *arXiv preprint arXiv:1506.02078* (2015).
[53] Kiwix. 2023. wikihow_en_maxi_2023-03.zim : WikiHow : Free Download, Borrow, and Streaming : Internet Archive. https://archive.org/details/wiki-how-en
[54] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances* (2025).
[55] Ryuto Koike, Liam Dugan, Masahiro Kaneko, Chris Callison-Burch, and Naoaki Okazaki. 2025. Machine Text Detectors are Membership Inference Attacks. *arXiv preprint arXiv:2510.19492* (2025).
[56] Kayvan Kousha and Mike Thelwall. 2026. How much are LLMs changing the language of academic papers after ChatGPT? A multi-database and full text analysis. *Scientometrics* (2026).
[57] Soonchan Kwon and Beakcheol Jang. 2025. A Comprehensive Survey of Fake Text Detection on Misinformation and LM-Generated Texts. *IEEE Access* (2025).
[58] Lucio La Cava, Luca Maria Aiello, and Andrea Tagarelli. 2026. Machines in the crowd? measuring the footprint of machine-generated text on reddit. *Online Social Networks and Media* (2026).
[59] Luke Leal. 2020. Hidden SEO Spam Link Injections on WordPress Sites. https://blog.sucuri.net/2020/11/hidden-seo-spam-link-injections-on-wordpress-sites.html Accessed 2026-04-28.
[60] Dirk Lewandowski, Sebastian Sünkler, and Nurce Yagci. 2021. The influence of search engine optimization on Google's results: A multi-dimensional approach for detecting SEO. In *Proceedings of the 13th ACM Web Science Conference 2021*.
[61] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
[62] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2025. Quantifying large language model usage in scientific papers. *Nature Human Behaviour* (2025).
[63] Runheng Liu, Heyan Huang, Xingchen Xiao, and Zhijing Wu. 2025. Zero-Shot Detection of LLM-Generated Text via Implicit Reward Model. In *Advances in Neural Information Processing Systems*.
[64] Dominik Macko, Aashish Anantha Ramakrishnan, Jason S Lucas, Robert Moro, Ivan Srba, Adaku Uchendu, and Dongwon Lee. 2026. Beyond Speculation: Measuring the Growing Presence of Large Language Model-Generated Texts in Multilingual Disinformation. *Computer* (2026).
[65] Ross A Malaga. 2008. Worst practices in search engine optimization. *Commun. ACM* (2008).
[66] Udi Manber et al. 1994. Finding similar files in a large file system. In *Usenix winter*.
[67] Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. RAIDAR: geneRative AI Detection viA Rewriting. In *International Conference on Learning Representations*.
[68] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*.
[69] Hoang-Quoc Nguyen-Son, Minh-Son Dao, and Koji Zettsu. 2026. SearchLLM: Detecting LLM Paraphrased Text by Measuring the Similarity with Regeneration of the Candidate Source via Search Engine. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*.
[70] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*.
[71] Petteri Nurmi, Musfira Khan, Zahra Safaei, Ngoc Thi Nguyen, Fatemeh Sarhaddi, Mika Tompuri, Henrik Nygren, Päivi Kinnunen, and Agustin Zuniga. 2026. Ai see what you did there–the prevalence of llm-generated answers in mooc responses. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V. 1*.
[72] Jack O'Connor, Jean-Philippe Aumasson, Samuel Neves, and Zooko Wilcox-O'Hearn. 2021. One function, fast everywhere. (2021).
[73] Pangram Labs. 2026. Pangram AI Detector. https://www.pangram.com/ Accessed October 4, 2026.
[74] Benjamin D Pesante, Cyril Mauffrey, and Joshua A Parry. 2024. Rise of the machines: the prevalence and disclosure of artificial intelligence–generated text in high-impact orthopaedic journals. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons* (2024).
[75] Pew Research Center. 2026. How Much of the Internet Is Written With AI? https://www.pewresearch.org/data-labs/2026/08/20/how-much-of-the-internet-is-written-with-ai/. Accessed October 13, 2026.
[76] Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. 2024. Ranking Manipulation for Conversational Search Engines. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
[77] ProjectDiscovery. 2025. projectdiscovery/wappalyzergo: A high performance go implementation of Wappalyzer Technology Detection Library. https://github.com/projectdiscovery/wappalyzergo
[78] Giovanni Puccetti, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. 2024. AI 'news' content farms are easy to make and hard to detect: A case study in Italian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
[79] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).
[80] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* (2020).
[81] Jenna Russell, Ben Glickenhaus, Katherine Thai, John Wieting, Mohit Iyyer, Max Spero, and Bradley Emi. 2026. How Much Is an AI Token Worth? Scaling Laws for Wild AI-Generated Web Text. https://arxiv.org/abs/2609.40295
[82] Jenna Russell, Marzena Karpinska, Destiny Akinode, James Zhou, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer. 2026. AI use in American newspapers is widespread, uneven, and rarely disclosed. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[83] Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.
[84] Giuseppe Russo, Manoel Horta Ribeiro, Tim Ruben Davidson, Veniamin Veselovsky, and Robert West. 2025. The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates. *Proceedings of the ACM on Human-Computer Interaction* (2025).
[85] McKenzie Sadeghi, Dimitris Dimitriadis, Virginia Padovese, Giulia Pozzi, Sara Badilini, Chiara Vercellone, Natalie Huet, Zack Fishman, Leonie Pfaller, and Natalie Adams. 2025. Tracking AI-enabled Misinformation: 1,254 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools. *NewsGuard* (2025).
[86] Kouichi Sakurai, Kaito Taguchi, and Yujie Gu. 2024. The Impact of Prompts on Zero-Shot Detection of AI-Generated Text. In *Proceedings of the IJCAI 2024 Workshop on Artificial Intelligence Safety (AISafety 2024)*.
[87] Sebastian Schultheiß and Dirk Lewandowski. 2021. "Outside the industry, nobody knows what we do" SEO as seen by search engine optimizers and content providers. *Journal of Documentation* (2021).
[88] Dongwon Shin and Sooel Son. 2026. LLMs Killed Q&A Stars? Analyzing the Impact of LLM-Generated Answers on an Online Q&A Platform. In *Proceedings of the ACM Web Conference 2026*.
[89] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. https://arxiv.org/abs/1908.09203
[90] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
[91] Nikita Spirin and Jiawei Han. 2012. Survey on web spam detection: principles and algorithms. *ACM SIGKDD explorations newsletter* (2012).
[92] Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
[93] Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.
[94] The EasyList authors. 2025. EasyList - Overview. (2025).
[95] Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*.
[96] Henry S Thompson. 2024. Improved methodology for longitudinal Web analytics using Common Crawl. In *Proceedings of the 16th ACM Web Science Conference*.
[97] Brian Tufts, Xuandong Zhao, and Lei Li. 2025. A Practical Examination of AI-Generated Text Detectors for Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*.
[98] Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.
[99] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2025. Prevalence and prevention of large language model use in crowd work. *Commun. ACM* (2025).
[100] Anthony Vinci. 2026. A.I. Is Lying to You. Here's How to Fight Back. https://www.nytimes.com/2026/10/06/opinion/ai-propaganda-information-spy.html. Published October 6, 2026.
[101] James Liyuan Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. 2024. RAFT: Realistic attacks to fool text detectors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
[102] Yi-Min Wang, Ming Ma, Yuan Niu, and Hao Chen. 2007. Spam double-funnel: Connecting web spammers with advertisers. In *Proceedings of the 16th international conference on World Wide Web*.
[103] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the twelfth language resources and evaluation conference*.
[104] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics* (2025).
[105] Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S Chao. 2024. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems* (2024).
[106] Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Ruth Petzold, William Yang Wang, and Wei Cheng. 2024. A Survey on Detection of LLMs-Generated Content. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
[107] Hongyeon Yu, Dongchan Kim, and Young-Bum Kim. 2026. Retrieval Collapses When AI Pollutes the Web. In *Proceedings of the ACM Web Conference 2026*.
[108] Xiao Yu, Yi Yu, Dongrui Liu, Kejiang Chen, Weiming Zhang, Nenghai Yu, and Jing Shao. 2025. Evobench: Towards real-world llm-generated text detection benchmarking for evolving large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*.
[109] Yunhao Zhang and Renée Gosline. 2023. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making* (2023).
[110] Yuexin Zhao, Siyi Tang, Hongyu Zhang, and Long Lyu. 2025. AI vs. human: A large-scale analysis of AI-generated fake reviews, human-generated fake reviews and authentic reviews. *Journal of Retailing and Consumer Services* (2025).
[111] Hongyi Zhou, Jin Zhu, Kai Ye, Ying Yang, Erhan Xu, and Chengchun Shi. 2026. Learn-to-Distance: Distance Learning for Detecting LLM-Generated Text. In *International Conference on Learning Representations*.
[112] Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. Beat LLMs at Their Own Game: Zero-Shot LLM-Generated Text Detection via Querying ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
[113] Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. 2025. Reliably Bounding False Positives: A Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

A Ethics

Crawling. In general, we only collect data that are already publicly available on the web, and do not attempt to collect any personally identifiable information. And, this work involves no human subjects and so does not require IRB review. All fetches honor robots.txt and apply a 1-minute delay between requests to the same domain. In our requests, we include a custom user-agent string that points to our project and contact information page for site operators to reach out with questions or concerns. To avoid stressing any hosts, for any given site, we stop requesting after three consecutive connection errors, and perform at most 640 page visits while crawling for samples. We also abide by Common Crawl's recommended connection limit and retry delay.

Generated baseline sites reachability. Our AI-generated baseline sites (§2.1) are kept unreachable from the public web (not indexed, not linked) to prevent them from entering future Common Crawl snapshots, search indices, or LLM training corpora.

B Generative AI Disclosure

Following ACM's policy on the use of generative AI tools, we disclose that the text in §5.3 and the appendices was originally drafted with the help of GPT-5.4, GPT-5.6 Sol, GPT-6.1 Sol, GPT-6 Astra, and Claude Opus 4.6 and 5.5, from our existing internal documents and experiment results. The initial drafts were extensively revised by the authors. The prompts in §D and §E were also initially generated with the help of AI tools. The authors also used AI coding agents to help write analysis code and to proofread or suggest edits to this paper. Besides the prompts, all sections have been reviewed, rewritten, and verified by the authors.

C Dolma Quality Filter

We use a relaxed version of the Dolma quality filter [90]. Our relaxations aim to preserve more valid documents that would otherwise be filtered out due to common boilerplate like footers on webpages.

C4 NoPunc [80]. removes lines not ending in punctuation '.' '?' '!' or '"'. Dolma also filters out documents with over half of their lines removed this way. Instead, we use NoPunc only as part of the text cleaning process, but not as a filtering criterion because we find it filters out many valid texts in practice due to boilerplate on webpages.

Gopher All [79]. Dolma uses the following heuristics to filter out documents with high repetition or low linguistic quality. They apply these heuristics to the original document, but we apply them to the NoPunc-cleaned document instead to preserve more valid documents despite noise from boilerplate on webpages.

- Fraction of characters in most common n-gram greater than threshold: bigram > 0.20, trigram > 0.18, 4-gram > 0.16.
- Fraction of characters in duplicate n-grams greater than threshold: 5-gram > 0.15, 6-gram > 0.14, 7-gram > 0.13, 8-gram > 0.12, 9-gram > 0.11, 10-gram > 0.10.
- Fewer than 50 or more than 100,000 words.
- Median word length < 3 or > 10.
- Symbol-to-word ratio > 0.10.
- Fraction of words with alphabetic characters < 0.80.
- Contains fewer than 2 required words: "the" "be" "to" "of" "and" "that" "have" "with".
- Fraction of lines starting with bullet > 0.90.
- Fraction of lines ending with ellipsis > 0.30.
- Fraction of duplicated lines > 0.30.
- Fraction of characters in duplicated lines > 0.30.

Dolma also removes documents "that contains a token or sequence of tokens repeating over 100 times" [90], but we do not use this criterion because it is insignificant, only filtering out 0.003% of characters for Dolma.

D Prompts for Generating Baseline Sites

We describe the prompt used to generate the Wix and B12 sites in §2.1. For each seed website, we gave this prompt to ChatGPT 4o in chatgpt.com, with `[website domain name]` replaced by the site's domain name. The prompt instructs the LLM to summarize and suggest the MGT-dominant website to be generated. ChatGPT was able to either conduct web search for describing the website or come up with descriptions based on the domain name if the website is not indexed. The prompt asks for a one-paragraph description of the site without its name, a similar site name, and 30 diverse blog post titles with descriptions, returned as JSON.

For the body-swap pages, a first call gives the LLM the site, page URL, and extracted main-body text, and asks for a 50- to 100-word paraphrased description that never reproduces 12 consecutive words of the source. A second call sees only that description and returns the HTML body, without mentioning the instructions. For 128 of the 328 polished sites, one prompt treats the page text as untrusted JSON and asks for a clean HTML fragment rewrite. Eight body-swap sites and the other 200 polished sites come from earlier runs whose exact prompts we did not keep.

E Website Categorization Prompts

Categorization of websites in §5.3 uses OpenAI's GPT-OSS-120B LLM. For each site, we randomly select up to 5 already-crawled sample pages, separate from the up to 3 live-fetched pages used for the other signals in §5.3. Each page makes a prompt fragment with its extracted main text and metadata (e.g., number of bytes or ads).

From these fragments, we assemble a prompt that asks the LLM to categorize the site using the prompt below. The LLM is free to either choose one of the approved categories, or propose a new category which the authors manually approve; the LLM may also assign zero or more tags to the site.

```
Task: classify the main purpose of subdomain 'SUBDOMAIN'.
Samples:
<sample_pages>
PAGE_FRAGMENTS
</sample_pages>
Approved categories:
- 'SLUG': use when USE_WHEN; exclude when EXCLUDE_WHEN...
Existing site tags:
- TAG...
Respond with 1 JSON object with quoted evidence,
reasoning, decision state, category, and tags.
```

We then parse the LLM's response as JSON, and reject any malformed responses.

F Monetization by Site Category

Table 6 breaks down the two audience-oriented monetization signals of §5.3 and the use of high-end ad platforms (Raptive, Sovrn, Mediavine, or PubMatic) by site category. Subscription/donation counts sites with donation or newsletter links (PayPal donations, Buy Me a Coffee, Patreon, Ko-fi, Substack, ConvertKit, Mailchimp, beehiiv, Ghost, Memberful, or Circle) on any crawled page, regardless of other signals; cultivation counts sites using Substack, ConvertKit, Mailchimp, beehiiv, or Ghost, regardless of other signals. In every category, both signals are less common on MGT-dominant sites. Averaging the per-category gaps, weighted by how many MGT-dominant sites each category has, gives gaps of −5.2 and −4.6 percentage points.

Table 6: Share of sites with subscription/donation, audience-cultivation, and high-end ad platform signals per site category, for MGT-dominant (MGT) and non-MGT-dominant (Non) sites.

	Sites		Subscr./don. (%)		Cultivation (%)		High-end ads (%)	
Category	MGT	Non	MGT	Non	MGT	Non	MGT	Non
Editorial	2,933	3,187	4.9	13.3	3.3	10.5	9.0	14.7
Service	3,578	4,021	2.1	5.6	1.9	5.0	0.1	0.6
Affiliate	985	882	2.8	6.7	1.4	5.4	8.0	14.5
Ecommerce	1,429	1,346	1.4	6.5	1.2	5.8	0.3	0.4
SaaS	343	434	5.0	6.9	4.7	6.5	0.3	0.0
Personal/org	227	2,811	5.7	10.8	4.8	8.8	3.1	7.5
Other	575	1,170	2.3	5.7	1.2	3.9	2.4	3.1
All	10,070	13,851	3.1	8.6	2.3	7.1	3.7	6.3

G Editorial Sites Without Clear Incentives

Across both datasets in §5.3, our methods detect no ads or affiliate links in 1,506 MGT-dominant publisher/editorial sites. Some of these sites may have undiscovered financial incentives, e.g., server-injected ads, scripts gated by consent [49] or login [51], or monetization that only occurs in select pages.

After inspecting select sites from *Search Dataset*, we observe mainly informational blogs or guides without clear monetization. Generic examples include wellness information sites (edailyworkout.com), tutorials on Word (learnword.io), and guides on tires (tirethink.com). Some niche topics include crystals (crystalhappiness.com), birds (birdswave.com), and gemstones (stongle.com). These may be from experimentation, hobby, or new projects pending monetization. We also find a potential demonstration site for a WordPress theme provider (artisticphotographyguide.com for generatepress.com), which acts as an example of a hidden incentive. Finally, sites like smallpetjournal.com appear defunct and contain mostly filler.

H Tech Stack Groups that Share AdSense IDs

Of the 270 tech-stack groups, only 36 groups (160 sites) also share an ID, email, or custom footer. In 19 groups (48 sites), members also share an AdSense ID or affiliate tag, which trace to 10 shared payees. Some payees use multiple tech stacks, e.g., 14 sites in the largest AdSense ID based cluster use 6 tech stacks. These sites all use WordPress, PHP, MySQL, Nginx, and Google AdSense, but differ in other detected technologies, including advertising services and analytics. Two groups of 6 and 2 Australian regional sites also share a single AdSense ID, e.g., whitsundays.guide.