Published by

British Journal of Contemporary Research

BEXFORD PUBLISHING

Algorithmic Poisoning: Quantifying the Infiltration of Synthetic Data in Open Public Knowledge Bases

Meta-Analysis **Volume 1 | Issue 1 | 2026**

Received: 01 June 2026 **Accepted: 06 June 2026**

Jhanna Mariel A. Merino[1] , Lorenzo Romero de la Cruz[1]

[1]Biliran Province State University, Naval, Biliran, Philippines

Correspondence: Lorenzo Romero de la Cruz, lorenzo.delacruz@bipsu.edu.ph

Abstract

The mass deployment of large language models (LLMs) since the late-2022 release of ChatGPT has produced a recursive feedback loop in which model-generated ("synthetic") text and code are scraped back into the corpora used to train subsequent models. Theoretical work warns that this loop can trigger "model collapse"—an irreversible loss of distributional tails and information integrity. This study synthesises and quantifies the rate of synthetic-data infiltration across three pillars of the open public knowledge commons—web crawls (Common Crawl), Wikipedia, and open code repositories—and evaluates how that infiltration correlates with degradation of information integrity. We conducted a structured review and meta-synthesis of peer-reviewed and government-sourced empirical studies, triangulating corpus-level detection estimates (excess-vocabulary maximum-likelihood models, perplexity and cross-perplexity detectors such as Binoculars, and proprietary classifiers such as GPTZero), provenance documentation analyses of Hugging Face dataset cards, and multi-way-parallelism audits of Common Crawl. Pre-March-2022 corpora served as calibrated false-positive-rate baselines. Convergent evidence shows synthetic infiltration rising sharply post-2022. Detectors flag over 5% of newly created English Wikipedia articles as AI-generated (lower bound 4.36%); at least 13.5% of 2024 biomedical abstracts show LLM excess-vocabulary signatures; 6.5–16.9% of AI-conference peer reviews are LLM-modified; and 57.1% of sentences in a Common Crawl-derived corpus are multi-way machine translations. Flagged content is systematically lower quality. Controlled experiments confirm model collapse under data replacement but show data accumulation bounds the damage. Synthetic infiltration of open knowledge bases is measurable, accelerating, and correlated with quality degradation, but is not yet catastrophic and is conditional on training methodology. Provenance standards, watermarking, and regulatory marking obligations are necessary but currently insufficient mitigations.

Keywords: Model Collapse, Synthetic Data, Data Provenance, Large Language Models, Information Integrity, Machine-Generated Text Detection, Common Crawl, Content Authenticity

1. Introduction

The open public knowledge commons—the web pages aggregated by Common Crawl, the collaboratively edited articles of Wikipedia, and the source code hosted in public repositories—constitutes both the principal training substrate for modern artificial intelligence and a shared epistemic infrastructure for humanity. Common Crawl alone releases monthly snapshots of web content scraped via automatic web crawling, dwarfing curated corpora such as English Wikipedia (≈5.6 TB) and BookCorpus (≈6 GB), and has been used to train the GPT series, BERT, and FastText (Luccioni & Viviano, 2021). Since the public release of ChatGPT in November 2022, generative models have begun to contribute content to these same repositories at scale. This produces a recursive feedback loop: models trained on web data generate text that is published to the web, scraped into the next crawl, and used to train the next model generation.

The principal theoretical hazard of this loop is model collapse. Shumailov, Shumaylov, Zhao, Papernot, Anderson, and Gal (2024), publishing in Nature, demonstrated that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. A parallel and intersecting concern is algorithmic poisoning—the degradation of the information value of a shared corpus as low-quality, homogenised, or fabricated synthetic content displaces organic human knowledge. Whereas model collapse is a property of the model, algorithmic poisoning is a property of the corpus; the two are coupled through the training–publication–scraping loop.

1.1 Background: Feedback Loops and Data Exhaustion

The urgency of the question is amplified by the fact that frontier models have nearly exhausted the available stock of high-quality human text (Borji, 2024). As clean human data becomes scarce relative to model appetite, the temptation—and necessity—to train on synthetic or web-scraped (and therefore increasingly synthetic-contaminated) data rises. The feedback loop is therefore not a hypothetical but a structural feature of the contemporary AI data economy.

1.2 Research Gap

Despite intense theoretical attention to model collapse, the empirical question—how much synthetic data has actually infiltrated open knowledge bases, and whether that infiltration measurably degrades integrity—has been addressed only in fragmented, domain-specific studies. No prior work integrates the Wikipedia, Common Crawl, scientific-corpus, and code-repository evidence into a unified, cross-source quantitative picture calibrated against the model-collapse theory. This paper closes that gap.

1.3 Research Question

What is the rate of synthetic-data infiltration within open-source text and code repositories, and how does it correlate with the degradation of information integrity?

1.4 Objectives

1. Quantify infiltration rates across Common Crawl, Hugging Face dataset documentation, Wikipedia, and adjacent high-quality corpora.
2. Characterise the statistical and linguistic signatures of LLM text and the reliability of detection.
3. Assess the correlation between synthetic prevalence and content-quality metrics.

4. Evaluate the model-collapse evidence base and its real-world applicability.
5. Appraise policy and provenance mitigations.

1.5 Significance

If the training substrate of AI is being contaminated by AI's own outputs, both the future trajectory of model capability and the integrity of human reference works are jeopardised. Quantifying the present rate is the necessary first step toward governance.

2. Literature Review

2.1 The Model Collapse Phenomenon

Shumailov et al. (2024) provided the canonical demonstration. Recursively fine-tuning an OPT-125M causal language model on its own generations from wikitext2, they observed that models progressively lose information about the true distribution, with tails (low-probability events) disappearing first and learned behaviours converging to a point estimate with very small variance. Their theoretical analysis showed that, for a discrete distribution, the tails begin to disappear as a result of the low probability of sampling them, and over time the support of the distribution shrinks, ultimately converging to a delta function. For the Gaussian case, the Wasserstein-2 distance from the original distribution diverges to infinity and the variance tends to zero with probability 1.

Their degenerate-text example is now widely cited: given an input about Perpendicular-style English church towers, the model's outputs degraded across generations, beginning with coherent if inaccurate prose and ending, by Generation 9, in repetitive lists of "tailed jackrabbits" variants. Quantitatively, recursive training without preserving original data produced large perplexity degradation, whereas preserving a small fraction of the original data led to only minor degradation.

Dohmatob, Feng, Yang, Charton, and Kempe (2024) formalised collapse through the lens of scaling laws, deriving a double scaling law and showing analytically that nucleus (top-p) sampling and temperature scaling truncate distributional tails, causing loss of scaling, shifted scaling with number of generations, the un-learning of skills, and grokking when mixing human and synthesised data. Critically, they also proved that mixing AI-generated data with even a small amount of clean data mitigates model collapse by introducing a grokking phenomenon.

A substantial counter-literature qualifies the catastrophist reading. Gerstgrasser et al. (2024) showed that collapse is conditional on a replace workflow: if synthetic data accumulates alongside a non-shrinking real-data anchor, test error no longer grows linearly with the number of model-fitting iterations, but instead has a finite and relatively small upper bound. Dey and Donoho (2024) established the universality of a $\pi^2/6$ bound on test risk for the accumulated workflow in classical linear regression, no matter how many generations of training with synthetic data take place. Borji (2024) independently replicated the kernel-density-estimation collapse, concluding the phenomenon is statistical and may be unavoidable under replacement.

2.2 Synthetic Data Generation, Diversity Loss, and Provenance

Guo, Shang, Vazirgiannis, and Clavel (2024) added the dimension of diversity loss, developing lexical, syntactic, and semantic diversity metrics and documenting a consistent decrease in the diversity of model

outputs through successive iterations, especially remarkable for tasks demanding high levels of creativity. This matters for corpus integrity independent of model collapse: even where capability is preserved, the variety of synthetic contributions to a knowledge base declines.

Hugging Face has become the dominant open repository for datasets and models. Yang, Gao, Xue, and Alfonseca (2024), analysing 24,065 dataset repositories, reported a 3.97% weekly growth rate and a doubling time of approximately 18 weeks, reaching 35,973 repositories by 23 May 2023. However, only 30.9% of repositories carried non-empty documentation, although those documented datasets accounted for 95.0% of download traffic. This documentation deficit is the core provenance problem: when synthetic data is uploaded without disclosure, downstream curators cannot exclude it.

2.3 AI Hallucination and Training-Data Degradation

The integrity hazard is compounded by hallucination—the generation of fluent but fabricated content, including fabricated citations. Wikimedia community documentation explicitly records instances of articles modified with LLMs that include hallucinated citations, and machine-translated articles created simultaneously across multiple language Wikipedias as cross-wiki automated-translation patterns (Wikimedia Foundation, 2025). Such fabrications, once embedded in a reference work, can be re-scraped and amplified.

2.4 Detection Methods and Statistical Signatures

Detection rests on the premise that LLM text exhibits low perplexity (next-token predictability) and low burstiness (sentence-to-sentence variance in length and complexity). DetectGPT exploited local probability curvature; the open-source Binoculars detector introduced normalised cross-perplexity, defined as the average cross-entropy between two models' outputs, and reports superior cross-domain performance. Corpus-level methods—Liang et al. (2024) and Kobak et al. (2025)—instead model word-frequency distributions as human/LLM mixtures or track excess vocabulary, avoiding reliance on ground-truth classifiers.

A persistent and important caveat is that detectors are unreliable at the individual-document level and on non-English and task-specific text. Quaremba, Black, Vrandecić, and Simperl (2025) found that on realistic Wikipedia editing tasks—paragraph writing, summarisation, and style transfer—training-based detectors achieve an average accuracy of 78%, while zero-shot detectors average 58%. Wikimedia's own editor guidance notes that humans distinguish LLM from human text at no better than random chance, and that marker words associated with ChatGPT in 2023 (such as "delve") declined sharply by 2025—evidence that signatures are a moving target as human and machine writing co-evolve.

2.5 Wikipedia, Common Crawl, and Code Contamination Studies

Brooks, Eggert, and Peskoff (2024), at Princeton University, provided the first systematic Wikipedia infiltration estimate. Thompson, Dhaliwal, Frisch, Domhan, and Federico (2024), at AWS AI Labs, audited Common Crawl for machine translation via multi-way parallelism. The composition of Common Crawl itself has long been scrutinised for undesirable content (Luccioni & Viviano, 2021), and audits of its derivatives (C4, OSCAR) surfaced machine-translated text and benchmark contamination well before the generative-AI boom. On code, large-scale industry analyses and academic GitHub-mining studies document measurable shifts in code composition and security following AI-assistant adoption.

3. Methodology and Research Design

This is a structured evidence-synthesis study triangulating three primary data sources, supplemented by adjacent high-quality corpora (scientific abstracts, peer reviews) that serve as sensitive sentinels for synthetic-text inflow.

3.1 Common Crawl (via AWS Open Data)

We reviewed corpus-composition audits of Common Crawl and its derivatives (C4, OSCAR, ccMatrix), focusing on the multi-way-parallelism method of Thompson et al. (2024). That method builds a multi-way-parallel corpus (MWccMatrix) from ccMatrix, embeds sentences with LASER, and infers machine generation from (a) the number of languages a sentence is translated into and (b) the declining quality of highly parallel sentences. The pre-2022 versus post-2022 framing required by the research question is operationalised via the GPT-3.5 (March 2022) and ChatGPT (November 2022) release boundaries, which the Wikipedia and scientific-corpus studies use as natural-experiment cut-points and false-positive-rate calibration baselines.

3.2 Hugging Face Model and Dataset Card Metadata

We reviewed large-scale documentation analyses of Hugging Face cards (Yang et al., 2024) to characterise dataset growth and the prevalence—and absence—of provenance documentation, including explicit synthetic-data declarations. Dataset cards adhere to a standard template covering data sources, collection methods, and considerations for use; their completeness is the operational proxy for declared provenance.

3.3 Wikipedia Edit-History Analysis

We synthesised studies applying detectors (GPTZero, Binoculars) and excess-vocabulary methods to Wikipedia new-page corpora. The key design, from Brooks et al. (2024), collected articles created in August 2024 in English, French, German, and Italian, and used a previously curated dataset of pages created prior to March 2022 as a pre-GPT-3.5 baseline. Detector thresholds were calibrated to induce a 1% false-positive rate on the baseline; the infiltration estimate is the post-2022 detection rate minus this calibrated baseline, yielding a lower bound. We additionally incorporated the WETBench task-specific reliability benchmark (Quaremba et al., 2025).

3.4 Analytical Framing

Infiltration rates are reported as detector-calibrated lower bounds, because false-negative rates (undetected synthetic content) are not estimated and are known to be higher for non-English and human-edited text. Integrity degradation is operationalised via quality proxies: references/footnotes per sentence, outgoing links per word, linguistic-diversity metrics, perplexity, code churn and duplication, and security-vulnerability density. The synthesis is explicitly correlational; causal claims are reserved for the controlled model-collapse experiments.

4. Results

4.1 Wikipedia Infiltration Rates and Quality Correlation

Brooks et al. (2024) estimated that approximately 4.36% of 2,909 English Wikipedia articles created in August 2024 contain significant AI-generated content (a 5.36% raw detection rate at a 1% calibrated

false-positive rate). Comparable analyses were performed for French (3,138 articles), German (3,907), and Italian (3,003), with lower estimated infiltration in the non-English languages. A subsample of English articles flagged by both GPTZero and Binoculars showed strong shared signal.

Critically, flagged articles were systematically lower quality on integrity proxies. AI-detected English articles averaged 0.667 footnotes per sentence versus 0.972 for all new articles, and 0.383 outgoing links per word versus 1.77—indicating both poorer sourcing and weaker integration into the Wikipedia link nexus. Manual inspection found self-promotional advertising (eight pages promoted businesses, restaurants, or websites) and politically polarising content (a further eight pages, including content related to non-neutral political portrayals and at least one user sockpuppet banned for fabricating Albanian historical material).

This is a lower bound: the authors note that higher false-negative rates for non-English text mean the actual amount of AI-generated content could be substantially higher. Detector reliability on realistic editing tasks is limited—Quaremba et al. (2025) found training-based detectors average only 78% accuracy and zero-shot detectors 58% on task-specific Wikipedia machine-generated text, which tends to resemble human-written text more closely. By 2025, English Wikipedia had updated its deletion policy (criterion G15) to permit speedy deletion of LLM-generated pages bearing clear evidentiary tells—leftover communication to the user (e.g., "as a large language model…", knowledge-cutoff disclaimers) or fabricated citations such as non-existent papers and unresolvable DOIs—an institutional acknowledgement of the scale of the problem.

4.2 Scientific Corpus Infiltration: A Sentinel for High-Quality Text Inflow

Scientific abstracts function as a high-signal sentinel because they are short, stylistically constrained, and dated. Kobak, González-Márquez, Horvát, and Lause (2025), analysing more than 15 million PubMed biomedical abstracts from 2010–2024, found via excess-vocabulary analysis that at least 13.5% of 2024 abstracts were processed with LLMs, reaching 40% for some subcorpora—an effect they characterised as unprecedented in quality and quantity, surpassing even the COVID-19 pandemic's imprint on scientific writing. The signature was a sharp post-ChatGPT rise in stylistic excess words: less-common words "delves" (frequency ratio r = 28.0), "underscores" (r = 13.8), and "showcasing" (r = 10.7), plus high-frequency "potential" (δ = 0.052), "findings" (δ = 0.041), and "crucial" (δ = 0.037). A complementary medRxiv corpus-level study of urology journals found the estimated AI-like-text proportion rose from 3.1% (2022) to 3.7% (2023) to a peak of 5.3% (2024) (Silent Author study, 2025).

Liang et al. (2024), applying a maximum-likelihood mixture model to AI-conference peer reviews, estimated that 6.5–16.9% of review text at ICLR 2024, NeurIPS 2023, CoRL 2023, and EMNLP 2023 could have been substantially modified by LLMs, versus virtually 0% in 2022–2023 reviews. The LLM-generated fraction was higher in reviews reporting lower confidence and submitted near deadlines. Latona, Ribeiro, Davidson, Veselovsky, and West (2024), in The AI Review Lottery (analysing 28,028 ICLR 2024 reviews with GPTZero), independently estimated that at least 15.8% of reviews (approximately 4,428) were written with AI assistance, with 49.4% of all submissions receiving at least one AI-assisted review—and found such reviews boosted paper scores and acceptance rates.

4.3 Common Crawl and Machine-Translation Contamination

Thompson et al. (2024), analysing a Common Crawl-derived corpus, found that of 6.38 billion sentences across 2.19 billion translation tuples, 3.63 billion (57.1%) were in multi-way-parallel (3+ language)

tuples—a strong signature of mass machine translation. Translation quality declined monotonically with the degree of parallelism, and low-resource languages were dominated by this machine-generated content. The authors found a selection bias whereby low-quality English content is translated en masse into many lower-resource languages via machine translation, likely for ad revenue, and warned this raises serious concerns about training models such as multilingual large language models. This machine-translation contamination predates the generative-AI boom but represents the first large-scale synthetic infiltration of web corpora and a structurally identical hazard.

4.4 Hugging Face Dataset Growth and the Documentation Deficit

Datasets on Hugging Face exhibit a 3.97% weekly growth rate, doubling every approximately 18 weeks (24,065 repositories by 16 March 2023, rising to 35,973 by 23 May 2023; Yang et al., 2024). Yet only 30.9% of repositories carried non-empty dataset cards—although those documented datasets accounted for 95.0% of download traffic. The documentation gap means that the provenance of the long tail of datasets, including whether they are synthetic, is largely undeclared, frustrating any curation-based mitigation of the feedback loop. Hugging Face itself hosts a rapidly growing population of explicitly synthetic datasets, illustrating that synthetic data is a first-class and expanding category of the repository.

4.5 Code Repository Contamination

AI coding-tool adoption is now near-universal: GitHub's 2024 survey of 2,000 enterprise developers across the United States, Brazil, Germany, and India found more than 97% reported having used AI coding tools at work (GitHub, 2024). The integrity consequences are measurable. GitClear's Coding on Copilot whitepaper (analysing 153 million changed lines authored January 2020–December 2023) reported that code churn—the percentage of lines reverted or updated within two weeks of authorship—was projected to double in 2024 relative to the 2021 pre-AI baseline, alongside rising code duplication and a shift from updated/deleted/moved code toward added/copy-pasted code (Harding, 2024).

Security analyses corroborate quality concerns. A large-scale CodeQL analysis of AI-attributed public GitHub files (Schreiber & Tippe, 2025) found that Python AI-generated code exhibited vulnerability rates of 16.18–18.50%, versus 8.66–8.99% for JavaScript and 2.50–7.14% for TypeScript, with critical weaknesses (SQL injection, OS command injection, hard-coded credentials) appearing in MITRE's 2024 Top 25; notably, 87.9% of files carried no CWE-mapped vulnerability, and 39% of collected AI-attributed files were documentation generation, an understudied maintainability vector. AI coding contamination of public repositories is now a structural feature of the commons.

4.6 Web-Wide Prevalence

Corpus-level commercial analyses, which must be read with detector-reliability caveats, indicate a tipping point. A non-peer-reviewed industry analysis by Graphite (graphite.io), examining approximately 65,000 English-language Common Crawl URLs (2020–May 2025) and classifying articles with the Surfer AI-content detector, estimated that in November 2024 the quantity of AI-generated articles published on the web surpassed human-written articles, with the share reaching approximately 52% (51.7%) as of May 2025 (Graphite, 2025). Because this estimate depends on an automated AI-content detector and because Common Crawl excludes much paywalled (and likely human-authored) content, it should be treated as indicative rather than definitive. Separately, the widely circulated forecast that as much as 90% of online content may be synthetically generated by 2026 originates from the 2022 Europol Innovation Lab foresight report Facing Reality? Law Enforcement and the Challenge of Deepfakes, where it is presented as an

expert estimate concerning synthetic media broadly (deepfakes and visual content), not as an empirical measurement of text corpora; it should not be cited as such (Europol, 2022). The contrast between the approximately 52% detector-estimated figure for text and the 90% expert forecast for synthetic media illustrates the gap between speculative foresight and corpus measurement.

4.7 Model Collapse: Experimental Magnitude and Real-World Bounds

Under controlled recursion with data replacement, Shumailov et al. (2024) quantified collapse: perplexity degraded substantially and outputs collapsed to repetitive degenerate text by generation 9, with low-probability tails vanishing first. Under data accumulation, however, Gerstgrasser et al. (2024) demonstrated empirically (across transformers, variational autoencoders, and diffusion models) and theoretically that error is bounded, and Dey and Donoho (2024) established the $\pi^2/6$ universality bound for linear regression. The real-world dynamic—data accumulating rather than wholesale replacement—therefore sits closer to the bounded regime, though a bounded-compute middle ground (accumulating data under a fixed compute budget) still showed real-test losses climbing faster, indicating residual risk.

Table 1. Synthetic-Data Infiltration-Rate Estimates Across Open Knowledge Sources

Source	Metric	Estimate	Study
English Wikipedia (new articles, Aug 2024)	Significant AI content	4.36% (lower bound; >5% raw)	Brooks et al. (2024)
PubMed biomedical abstracts (2024)	LLM-processed	≥13.5% (up to 40% subcorpora)	Kobak et al. (2025)
Urology journal abstracts	AI-like text (2022→2024)	3.1% → 5.3%	medRxiv (2025)
AI-conference peer reviews (2023–24)	LLM-modified text	6.5–16.9%	Liang et al. (2024)
ICLR 2024 peer reviews	AI-assisted (GPTZero)	≥15.8% (~4,428 reviews)	Latona et al. (2024)
Common Crawl-derived sentences	Multi-way machine-translated	57.1%	Thompson et al. (2024)
Web articles (Common Crawl URLs)	AI-generated (parity point)	~50% from Nov 2024	Graphite (2025)

Hugging Face datasets	With any documentation	30.9%	Yang et al. (2024)
Enterprise developers	Have used AI coding tools	>97%	GitHub (2024)

Table 2. Integrity-Degradation Proxies in AI-Flagged Wikipedia Content (English)

Metric	AI-detected articles	All new articles
Footnotes per sentence	0.667	0.972
Outgoing links per word	0.383	1.77

Source: Brooks, Eggert, and Peskoff (2024).

Table 3. Model-Collapse Evidence: Replacement Versus Accumulation

Workflow	Outcome	Quantitative Result	Study
Replace (synthetic only)	Collapse	Substantial perplexity degradation; degenerate output by gen 9	Shumailov et al. (2024)
Replace (theory)	Collapse	Test error grows linearly with generations	Dohmatob et al. (2024)
Accumulate (real + synthetic)	Bounded	Error bounded ($\pi^2/6$ for linear regression)	Gerstgrasser et al. (2024); Dey & Donoho (2024)
Recursive fine-tuning	Diversity loss	Monotonic decline in lexical/syntactic/semantic diversity	Guo et al. (2024)

Figure 1 (Description)

A time series of synthetic-content prevalence against the November 2022 ChatGPT release would show near-zero detection in all corpora through early 2022, an inflection at the GPT-3.5/ChatGPT boundary, and divergent trajectories by 2024: scientific abstracts (~13.5%) and peer reviews (~16%) rising fastest among high-quality corpora, Wikipedia new articles at ~5%, and web-wide article prevalence reaching ~50%.

Overlaid, the "delve" frequency curve would peak in early 2024 and decline thereafter—visually encoding the co-evolution that erodes detector reliability.

Figure 2 (Description)

A scatter of Wikipedia article quality (footnotes per sentence; outgoing links per word) against AI-detection status would show AI-flagged articles clustering in the low-sourcing, low-integration quadrant, visualising the infiltration–integrity correlation of Table 2.

5. Discussion

5.1 Convergent Interpretation

The principal finding, robust across independent methodologies (calibrated black-box detectors, cross-perplexity, excess-vocabulary mixture models, multi-way parallelism) and across distinct corpora (Wikipedia, PubMed, peer review, Common Crawl, GitHub), is that synthetic-data infiltration is real, measurable, accelerating post-2022, and correlated with reduced quality. The Wikipedia footnote and link deficits (Table 2), the linguistic-diversity decline documented by Guo et al. (2024), and the rising code churn and vulnerability density all point in the same direction: synthetic content is, on average, less richly sourced, less interconnected, and less diverse than the organic content it displaces. This convergence is the study's central contribution—prior work established each datapoint in isolation; integrated, they describe a coherent, cross-domain phenomenon.

5.2 Comparison with Prior Work and the Catastrophist Debate

The catastrophist reading—that the web is becoming a Habsburg information ecosystem doomed to collapse—is not warranted by current evidence. Three lines of evidence bound the threat: (1) the Gerstgrasser et al. (2024) accumulation result and the $\pi^2/6$ theory show that the realistic data dynamic (accumulation, not replacement) bounds error; (2) Dohmatob et al.'s (2024) grokking result shows even small clean-data fractions mitigate collapse; and (3) the empirical plateau of web-wide AI prevalence around 50%, rather than runaway exponential growth, is consistent with an equilibrium rather than a collapse spiral. The most defensible synthesis is that model collapse is a genuine but conditional hazard, while algorithmic poisoning of human reference works is the more immediate, already-realised integrity concern. The 4.36% of new Wikipedia articles and the systematic quality deficits matter today, regardless of whether frontier models ever collapse.

5.3 Theoretical Implications

The results suggest the field should reframe model collapse from a binary fate to a continuous risk function parameterised by the real-to-synthetic ratio, the replace-versus-accumulate workflow, and the compute budget. The corpus-level algorithmic-poisoning lens complements this by treating the commons itself as the degrading object, with its own quality metrics (sourcing, diversity, interconnection) that degrade before and independently of any downstream model collapse.

5.4 The Detection-Reliability Caveat

A central methodological limitation colours all infiltration estimates: detector unreliability. Perplexity- and burstiness-based methods can be defeated by paraphrasing and are unreliable in high-stakes individual cases; task-specific machine-generated text evades detection (Quaremba et al.'s 58% zero-shot accuracy);

and human–machine co-evolution erodes signatures—the very "delve" marker that anchored the 2023–24 studies had faded by 2025. Consequently, every infiltration figure in this study is best read as a lower bound with a widening confidence interval over time.

6. Policy Implications

6.1 Data-Provenance Standards and C2PA

The U.S. National Institute of Standards and Technology's report NIST AI 100-4, Reducing Risks Posed by Synthetic Content (Chandra et al., 2024), catalogues digital watermarking and metadata-recording approaches and endorses the Coalition for Content Provenance and Authenticity (C2PA) specification, which attaches cryptographically signed manifests declaring content creation and editing history. NIST cautions, however, that adoption is nascent, that widespread use will also depend both on users opting in and on users and platforms not stripping the metadata, and that even watermarks that are robust to specific classes of perturbations can be vulnerable to adversarial attacks. Recent security analyses confirm C2PA manifests remain vulnerable to re-signing and metadata-washing attacks.

6.2 Watermarking for Curation, Not Just Deception Prevention

Sander, Fernandez, Durmus, Douze, and Furon (2024), published at NeurIPS, showed that watermarked text is radioactive—detectable in downstream models fine-tuned on it with high confidence ($p < 10^{-5}$) even when only 5% of fine-tuning data is watermarked. This reframes watermarking's value: beyond flagging deception, it enables curation—future data collectors can identify and exclude synthetic content from training sets, directly attacking the feedback loop. Robustness limits remain: image watermarks often fail to survive training (lose radioactivity), and watermark robustness and radioactivity can be at odds in federated settings.

6.3 Regulatory Frameworks

The European Union's Artificial Intelligence Act, Article 50, requires that providers of generative systems ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated, with obligations effective, interoperable, robust, and reliable (European Commission, 2024). These transparency obligations become enforceable on 2 August 2026, backed by penalties of up to €15 million or 3% of global annual turnover. The EU AI Office's draft Code of Practice on Transparency of AI-Generated Content (first draft published 17 December 2025) specifies a multi-layered approach—signed metadata plus watermarking plus optional fingerprinting—and explicitly states that no single marking technique currently meets the Article 50(2) requirements. A carve-out exempts standard editing (grammar, spellchecking) but not substantive alteration (AI translation, summarisation, object removal).

6.4 Staged Recommendations

1. Immediate (0–12 months): Repository operators (Hugging Face, GitHub, Wikimedia) should mandate machine-readable provenance fields and synthetic-data declarations on uploads; the current 30.9% documentation rate is the binding constraint. Wikimedia should extend criterion-G15-style review and deploy open-source detectors (Binoculars) as triage, not adjudication.

2. Near-term (12–24 months): Model developers should adopt accumulation-preserving training with a retained, audited real-data anchor and document the real-to-synthetic ratio—the single most evidence-supported collapse mitigation. Align generation pipelines with C2PA and Article 50 ahead of the August 2026 deadline.
3. Structural (24+ months): Governments (NIST, the EU AI Office, OECD) should fund independent, longitudinal corpus-monitoring of Common Crawl and reference works, and standardise cross-detector benchmarks.

Benchmarks that would change these recommendations: If independent monitoring shows web-wide AI prevalence breaking decisively above its ~50% plateau toward 70–80%, or if Wikipedia infiltration exceeds ~15% with continued quality deficits, the posture should escalate from voluntary provenance to mandatory pre-ingestion filtering. Conversely, if watermark adoption surpasses ~50% of generated content with durable radioactivity, curation-based mitigation becomes viable and detection-based triage can be de-emphasised.

7. Limitations

All infiltration figures are detector-calibrated lower bounds with non-trivial and growing false-negative rates, especially for non-English, short, and human-edited text. Several primary sources are arXiv preprints rather than fully peer-reviewed journal articles, though many are peer-reviewed conference papers (ICML, NAACL, NeurIPS, ICLR) or appear in government repositories. Two figures rest on lower-quality evidence and are flagged as such at point of use: the web-wide prevalence estimate (~52%) derives from a non-peer-reviewed industry analysis using an automated AI-content detector, and the 90% figure is an expert foresight estimate concerning synthetic media rather than a measurement of text. The Hugging Face documentation figures date to 2023 and predate the most recent synthetic-data surge. The synthesis is correlational: the association between AI-detection and low quality (Table 2) does not establish that synthetic generation causes degradation rather than that low-effort contributors use both AI and poor sourcing. Finally, detection thresholds, model versions, and human–machine co-evolution all drift, limiting the comparability of estimates across study dates.

8. Future Research Directions

Priorities include: (1) longitudinal token-distribution studies directly comparing pre-2022 and post-2022 Common Crawl snapshots via the AWS Open Data archive, quantifying excess-vocabulary drift at web scale; (2) standardised, adversarially robust cross-detector benchmarks that report calibrated false-negative as well as false-positive rates on task-specific and multilingual text; (3) causal, frontier-scale model-collapse experiments that vary the real-to-synthetic ratio and replace-versus-accumulate workflow under realistic compute budgets; (4) provenance-adoption tracking measuring the fraction of generated content carrying durable C2PA or watermark signals; and (5) integrity-metric standardisation for knowledge bases (sourcing density, interconnection, diversity) to enable consistent cross-corpus monitoring.

9. Conclusion

Synthetic data has measurably infiltrated the open knowledge commons—on the order of 5% of new English Wikipedia articles, at least 13.5% of biomedical scientific abstracts, 6.5–16.9% of AI-conference peer reviews, and a 57.1% majority of multilingual web translations—and this infiltration began accelerating sharply at the November 2022 ChatGPT inflection. The infiltration correlates with concrete

declines in information integrity: poorer sourcing and interconnection in flagged Wikipedia articles, reduced linguistic diversity under recursive training, and rising churn and vulnerability density in code. Model collapse is empirically real but conditional on training methodology: it occurs under data replacement and is bounded under the realistic regime of data accumulation with a retained real-data anchor. The more pressing near-term risk is therefore not the sudden collapse of frontier models but the steady erosion of the quality of human reference works as they absorb undeclared, lower-quality synthetic content. Provenance standards (C2PA), curation-enabling watermarking, and the EU AI Act's Article 50 marking obligations form the necessary mitigation architecture, but all remain immature—technically fragile, voluntarily adopted, and not yet enforced. Closing the gap between this fragmentary defensive posture and the measured, accelerating rate of infiltration is the central governance challenge for the integrity of open knowledge in the generative-AI era.

References

1. Borji, A. (2024). A note on Shumailov et al. (2024): "AI models collapse when trained on recursively generated data." arXiv. https://doi.org/10.48550/arXiv.2410.12954

2. Brooks, C., Eggert, S., & Peskoff, D. (2024). The rise of AI-generated content in Wikipedia. In Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia (WikiNLP 2024) (pp. 67–79). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wikinlp-1.12

3. Chandra, B., Dunietz, J., Roberts, K., Lee, Y., Fontana, P., & Awad, G. (2024). Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency (NIST AI 100-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-4

4. Dey, A., & Donoho, D. L. (2024). Universality of the $\pi^2/6$ pathway in avoiding model collapse. arXiv. https://doi.org/10.48550/arXiv.2410.22812

5. Dohmatob, E., Feng, Y., Yang, P., Charton, F., & Kempe, J. (2024). A tale of tails: Model collapse as a change of scaling laws. In Proceedings of the 41st International Conference on Machine Learning (PMLR Vol. 235). https://doi.org/10.48550/arXiv.2402.07043

6. European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 50. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

7. European Commission, Shaping Europe's Digital Future. (2025). Code of Practice on marking and labelling of AI-generated content. https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content

8. Europol. (2022). Facing reality? Law enforcement and the challenge of deepfakes (Europol Innovation Lab observatory report). Publications Office of the European Union. https://doi.org/10.2813/158794

9. Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Korbak, T., Sleight, H., Agrawal, R., Hughes, J., Pai, D. B., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., & Koyejo, S. (2024). Is model

collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. arXiv. https://doi.org/10.48550/arXiv.2404.01413

10. GitHub. (2024). Survey: The AI wave continues to grow on software development teams. GitHub Blog. https://github.blog/news-insights/research/survey-ai-wave-grows/

11. Graphite. (2025). More articles are now created by AI than humans. Graphite Five Percent Research. https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans

12. Guo, Y., Shang, G., Vazirgiannis, M., & Clavel, C. (2024). The curious decline of linguistic diversity: Training language models on synthetic text. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 3589–3604). https://doi.org/10.18653/v1/2024.findings-naacl.228

13. Harding, W. (2024). Coding on Copilot: 2023 data shows downward pressure on code quality. GitClear. https://www.gitclear.com/coding_on_copilot_data_shows_ais_downward_pressure_on_code_quality

14. Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Science Advances, 11(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813

15. Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024). The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. arXiv. https://doi.org/10.48550/arXiv.2405.02150

16. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In Proceedings of the 41st International Conference on Machine Learning (PMLR Vol. 235, pp. 29575–29620). https://doi.org/10.48550/arXiv.2403.07183

17. Luccioni, A. S., & Viviano, J. D. (2021). What's in the box? A preliminary analysis of undesirable content in the Common Crawl corpus. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Vol. 2: Short Papers) (pp. 182–189). https://doi.org/10.48550/arXiv.2105.02732

18. Quaremba, G., Black, J., Vrandecić, D., & Simperl, E. (2025). WETBench: A benchmark for detecting task-specific machine-generated text on Wikipedia. arXiv. https://doi.org/10.48550/arXiv.2507.03373

19. Sander, T., Fernandez, P., Durmus, A. O., Douze, M., & Furon, T. (2024). Watermarking makes language models radioactive. Advances in Neural Information Processing Systems 38 (NeurIPS 2024). https://doi.org/10.48550/arXiv.2402.14904

20. Schreiber, R., & Tippe, U. (2025). Security vulnerabilities in AI-generated code: A large-scale analysis of public GitHub repositories. arXiv. https://doi.org/10.48550/arXiv.2510.26103

21. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

22. Thompson, B., Dhaliwal, M. P., Frisch, P., Domhan, T., & Federico, M. (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 (pp. 1763–1775). https://doi.org/10.18653/v1/2024.findings-acl.103

23. Wikimedia Foundation. (2025). Investigate automated detection methods for LLM-generated and machine-translated content (Phabricator Task T406818). https://phabricator.wikimedia.org/T406818

24. Yang, X., Gao, J., Xue, W., & Alfonseca, E. (2024). Navigating dataset documentations in AI: A large-scale analysis of dataset cards on Hugging Face. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). https://doi.org/10.48550/arXiv.2401.13822