

Noise Floor

The output without an author — an evidence-led study of AI-generated content at scale, the marks meant to identify it, the rules that require them, and who checks the volume

*Output volume is not the same as verified human attention. The systems that receive the output cannot always tell who or what **sent it**.*

REPORT • V1.0 • PUBLIC RELEASE • EVIDENCE CUTOFF 2026-10-18

About this report

Shadow is an independent research series by Foresight Future Institute. This issue examines AI-generated content at scale: how much of it there is, the marks meant to identify it, the rules that require those marks, who carries the cost of checking it, and what it does to learning.

The report is evidence-led. Every factual statement is tied to a numbered source in the References. Where an organisation reports its own figures, the text says so. Chapter 7 lists what the evidence does not support. A summary of the method is at the end.

Contents

	Executive summary: five findings	3
	Disclosure and research notice	3
01	The floor rises	4
02	Who wrote this?	8
03	Marks that fade	12
04	Who must mark what	15
05	Who checks the volume	22
06	Assisted scores, unaided skills	25
07	What is not established	29
	Method	31
	References	32

— FIGURES

F1	Signs of AI authorship by web domain, 2021–2026	5
F2	Music: reported upload share and stream share	6
F3	arXiv monthly submissions	7
F4	Counting, screening, judging	9
F5	1.5 million registrations, 17,000 owners	12
F6	Providers, one table	13
F7	Who must mark what	19
F8	Different samples, periods and units	24
F9	Practice versus exam	26
F10	UK undergraduates and AI, 2024–2026	27

Executive summary: five findings

1. **Scale.** In Pew's July 2026 sample of 10,000 English-language pages, 10% showed significant signs of AI authorship. In the subset of pages dated after November 2022, a non-random minority of the crawl, the share exceeded one third. On .com, Pew's six-month aggregates rose from about 1% to 9.35% in the period labelled January 2026.¹
2. **Measurement.** Estimating how much of a corpus is AI-written and making an adverse decision about one person need different evidentiary treatment. A court ruling, several universities' decisions and a major conference's procedures each reflect this in different ways; their decisions do not establish one universal rule.^{2,3,4,5,6,7,8}
3. **Marking.** A detected mark can be evidence that an output was generated or processed by a tool covered by a given scheme. Coverage, access to checking and robustness to editing vary by implementation. Selected published evaluations report that editing weakens text marks. (Figure 6)^{9,10}
4. **Volume.** arXiv reported roughly twice as many submissions in September 2026 as in September 2024, and limited each submitter to two new submissions a month.¹¹
5. **Learning.** In two controlled studies, AI assistance improved scores on assisted tasks but did not establish a corresponding advantage on unassisted outcomes. In the mathematics study, unassisted exam scores fell after unrestricted access, which the authors report a safeguarded tutor largely mitigated. The outcomes and interventions differed.^{12,13,14}

Disclosure and research notice

This issue is written in a personal capacity by contributors to the Foresight Future Institute. It does not represent any contributor's employer. No case examined describes an engagement of any contributor's employer or its clients. Models from more than one developer were used for drafting, verification and review; material about each AI provider was checked by a model from a different provider.

RESEARCH NOTICE AND METHODOLOGY DISCLAIMER

1. **Scope.** This study is conducted in a personal capacity by contributors to the Foresight Future Institute. It is independent editorial research into AI-generated content, its marking and its verification. It does not reflect the views of any employer or official body.
2. **What "not located" means.** A statement that a document, rule or feature was not located means that it was not retrieved in the searches made for this issue, up to the close of the record. It is not a statement that it does not exist, and not a statement about the internal workings of any body.
3. **No ranking, no evaluation.** The report does not rank countries, providers, institutions or platforms, and does not evaluate their performance. It does not evaluate the implementation or outcomes of the measures it describes.
4. **Legal and policy summaries.** Summaries of laws, guidance and policies are limited to the instruments and dates specified. They are not an exhaustive statement of applicable law, and not an assessment of anyone's compliance.

5. **Reports, classifications and findings.** Figures and classifications published by organisations are reported as theirs and attributed in the sentence. A detector score, a vendor classification or a platform's own label is not a finding of misconduct. Decisions of courts and official bodies are identified as such and reported within their stated scope.
6. **Named bodies.** Authorities, institutions and companies are named only where a public document concerns them, and each statement reports what that document says. Naming does not imply endorsement or criticism.
7. **AI assistance.** The text was prepared with AI assistance and checked as described in the Method. Errors are possible. Every factual statement carries a numbered reference, so that a reader can check it.
8. **Not advice, no liability.** Nothing here is legal, financial or professional advice. To the extent permitted by applicable law, the authors and the Foresight Future Institute accept no liability for decisions that third parties take on the basis of this report.
9. **Sources and currency.** Links lead to documents of third parties, which can change or be withdrawn. The record closes on 18 October 2026, 23:59 (UTC+4). Later documents are not absorbed into this edition.
10. **Corrections.** Any body or reader may send a verifiable document to info@foresightflow.org. Corrections are dated and published, and the earlier text stays visible.

01

The floor rises

Opening: ninety thousand songs a day

In July 2026, Deezer, the Paris-based music service, reported that fully AI-generated tracks had exceeded half of its daily uploads at a June peak of roughly 90,000 tracks a day¹⁵. Its reported figure for January 2025 was about 10,000¹⁶.

Deezer also reported a much smaller share of streams for fully AI-generated music, between one and three percent, and classified up to 85% of those streams in 2025 as fraudulent¹⁵. These are different measures over different periods. They do not establish a matched-cohort listening rate, or the size of a human audience.

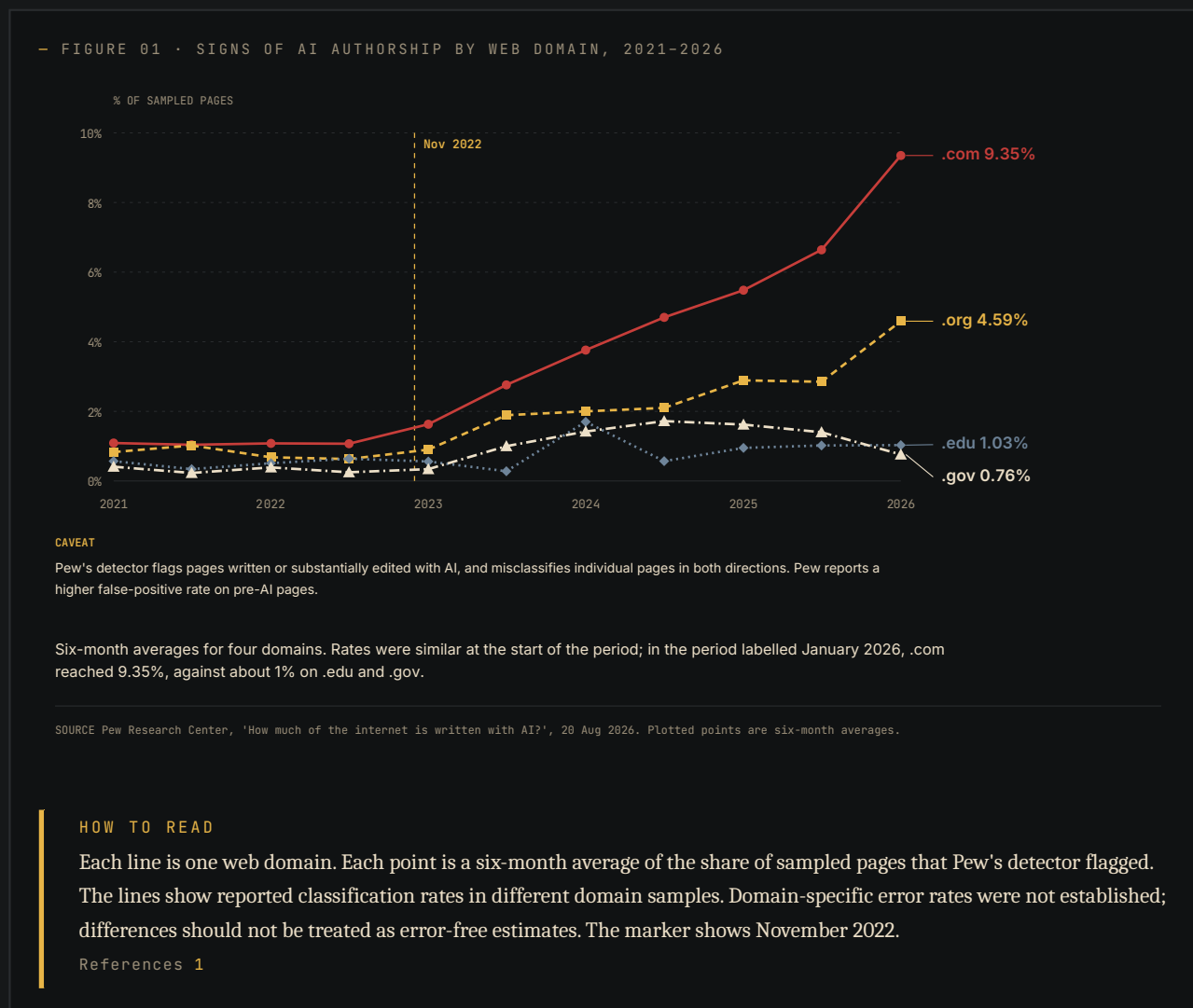
That distinction runs through this issue. "Slop" is usually described as a problem of quality. The evidence here is about something else. Output volume is not the same as verified human attention, and the systems receiving the output cannot always tell who or what sent it.

1.1 The web, measured

In August 2026 the Pew Research Center published an analysis of roughly 490,000 English-language pages sampled from the Common Crawl archive between January 2021 and July 2026. Each page was scored with an open-weight detection model at a stated threshold¹.

In a random July 2026 sample of 10,000 pages, old pages included, about 10% showed significant signs of AI authorship¹. Among pages carrying a publication date after November 2022, the share was over one third¹. Only about 10–15% of crawl pages carry usable publication-date information, and the post-November-2022 analysis uses a subset of those pages. That subset is not a random sample of recent pages, and its rate should not be extended to all of them.

Pew's chart compares four domains (.com, .org, .edu and .gov) using six-month averages. Their measured rates were similar at the start of the period and diverged over it. In the period labelled January 2026 the .com average reached 9.35%, against about 1% on .edu and .gov (Figure 1)¹.



Pew is explicit about what this does and does not show. Its threshold captures pages *substantially edited* with AI as well as those generated outright. Detectors misclassify individual pages in both directions¹. Its methodology notes a higher false-positive rate on pages written before AI tools were common¹. A higher error rate on early pages raises the measured baseline; on its own it does not produce the later increase. Each period's error rate should be read on its own terms¹. The study publishes its sampling and classification method. It does not issue a verdict on any individual page, a distinction Chapter 2 returns to.

1.2 Music: uploads and listeners

Other music services have announced their own content policies and enforcement measures. In September 2025 Spotify said it had removed more than 75 million "spammy" tracks over the preceding twelve months; that figure is not a count of AI-generated tracks^{17,18}. In January 2026 Bandcamp barred music generated wholly or in substantial part by AI^{17,18}.

Upload share and stream share measure different activities. Deezer's figures show a wide gap between them (Figure 2)^{15,16}. They do not establish a listening rate for a matched set of tracks, or the size of any human audience¹⁵.



1.3 Video and feeds

Several large platforms have launched feeds built around AI-generated video, and video platforms have revised their policies on repetitive and mass-produced content^{19,20,21}. How much of what a viewer sees is synthetic is harder to establish. The video-editing company Kapwing followed one new account's first 500 short videos. It classified 21% as "AI slop" and 33% as "brainrot" by its own definitions, categories that may overlap²². That single-account case study does not establish population-wide exposure.

1.4 News that looks like news

NewsGuard, which rates news sources, reports a count of "AI content farms". These are sites that publish substantially AI-generated material with little human oversight, present themselves as news outlets and do not adequately disclose it. The tracker, retrieved in October 2026, displayed an update dated 23 June 2026 and a count of 3,749 such sites in 16 languages, Arabic among them²³. The criteria are published. The count is NewsGuard's own inventory, not a census of news sites. NewsGuard's detection work involves Pangram, a detection company whose figures appear in Chapter 2. Reports that share detection infrastructure should not be treated as independent validation of that infrastructure^{23,24}.

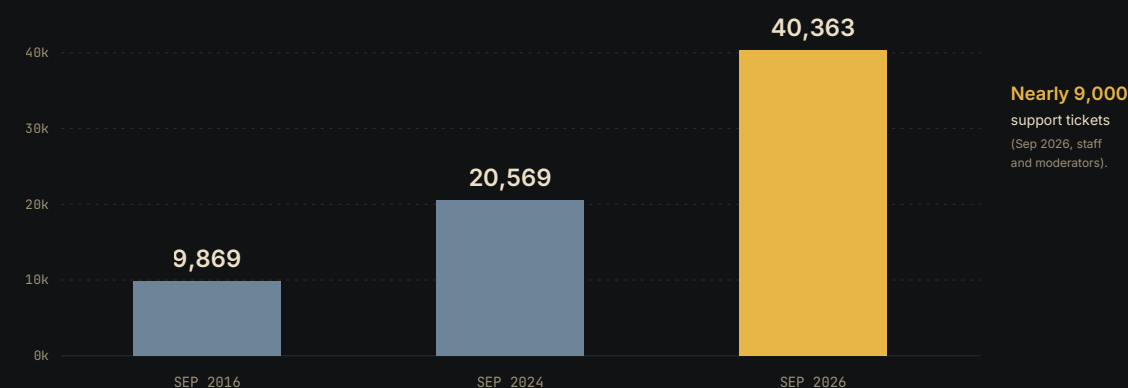
Not every investigator relies on detectors. Jean-Marc Manach, a French journalist at Next.ink who investigates AI-generated French-language "news" sites, says: "I've never used automatic AI detectors." He identifies sites through their metadata and other corroborating signs²⁵.

1.5 Science

On 1 October 2026 arXiv, the preprint server through which much of physics, mathematics and computer science first circulates, limited each submitter to two new submissions a month¹¹. Its figures explain the pressure. In September 2016 arXiv received 9,869 submissions; in September 2024, 20,569; in September 2026, 40,363 (Figure 3)¹¹. Over 2024–26, submissions to the artificial-intelligence category grew more than sixfold, against roughly twofold elsewhere¹¹. Moderators report thin papers, single studies split across several papers, and a "marked increase in dense, AI-written papers"¹¹.

— FIGURE 03 · ARXIV MONTHLY SUBMISSIONS

SEPTEMBER SUBMISSIONS



cs.AI submissions: more than 6x in 2024–2026; other categories roughly doubled.

Roughly twice as many submissions in September 2026 as in September 2024. From 1 October 2026 each submitter may send two new papers a month; rejected ones count.

arXiv attributes a disproportionate share of moderator time to a relatively small proportion of authors.

SOURCE arXiv blog, 'Updated rate limit policy', 1 Oct 2026.

HOW TO READ

Each bar counts the submissions arXiv received in one September. The 2026 bar is roughly double the 2024 bar. The new limit applies to each submitter, and rejected submissions count towards it.

References 11

arXiv's own account describes a concentration of workload. The chair of its editorial advisory council said a relatively small proportion of authors submit many low-quality papers and take up a disproportionate share of moderators' time¹¹. The limit nevertheless applies to every submitter. It counts rejected submissions, and it applies per submitter, not per co-author¹¹.

In medicine, an audit published in *The Lancet* found fabricated references in 1 in 277 papers in the corpus it examined from the first seven weeks of 2026²⁶. Fabricated references are not proof that a paper was written with AI. The audit does not establish which checks were attempted on these papers or why they failed.

1.6 The archive and the models

Common Crawl, the archive Pew sampled, is also a documented source of training text for language models. A 2024 Mozilla Foundation study of 47 models released between 2019 and 2023 found that at least 64% had been trained on Common Crawl²⁷. In a policy submission, Common Crawl itself has estimated that its data makes up 70–90% of training tokens²⁸. None of this establishes that the pages Pew measured entered any particular training corpus.

Research on models learning from model output gives conditional answers. A 2024 paper in *Nature* found irreversible defects under the recursive-training procedure it studied, in which each model is trained indiscriminately on the previous model's output²⁹. Other work finds materially different outcomes when synthetic data accumulates alongside the original real data instead of replacing it³⁰. Whether the web content Pew measured is degrading today's models has not been established, and this issue does not claim it.

02

Who wrote this?

Authorship and traffic measurements are method- and population-dependent. Validated aggregate estimates can be informative even when an individual classification is insufficient for disciplinary or other adverse action.

Opening: a court reviews one record

In November 2024 a student at Adelphi University in New York submitted an essay. His professor reported that Turnitin's AI-writing detector gave it a score of 100%, and that the student had said he used Grammarly. The student denied using AI to write the essay, and results from two other detectors, submitted on his behalf, reported a 0% likelihood. The university found a violation of its academic-integrity rules and required him to attend a plagiarism workshop. Failing to complete it would bring a hold on his registration⁶.

In January 2026 the Supreme Court of New York, Nassau County, annulled the determination and ordered it expunged from his record. The court found the decision "without valid basis and devoid of reason"⁶. Its grounds were specific to the case:

- the university had not considered the contrary evidence;
- the professor later wrote that he had expected the integrity office to assess the file independently, which the court treated as undermining the allegation;
- the student had been denied an adviser of his choice;
- the officer who made the determination also decided the appeal.

The court made no general finding on whether AI detectors work⁶.

Matter of Newby v Adelphi University reviewed the adequacy of one evidentiary record and one disciplinary process; the court did not estimate the detector's error rate. The case illustrates the problem this chapter is about: a score produced by one method was treated as settling a question about one person.

2.1 Counting versus judging

There are two kinds of question. "How much of this corpus was written with AI?" is a question about a population. "Did this person write this essay?" is a question about an individual. They need different evidence.

Estimating a share across a corpus requires a method validated for that population. A larger sample reduces sampling uncertainty, but it does not by itself remove systematic classification bias. A 2024 paper at ICML, a major machine-learning conference, set out a method for estimating the share of AI-modified text in a corpus and drew the distinction between aggregate trends and individual detection⁸. Pew publishes its sampling and classification method and acknowledges its classification limits¹.

The individual question is different again. Another ICML 2024 paper reported strong detection at a very low false-positive rate on the distributions it tested³¹. A document-level detector score is evidence produced by a particular method. What it means depends on calibration, on the conditions in which the detector is used and on the question being decided. Figure 4 shows how the same kind of instrument is used for counting, for screening and for judging.

— FIGURE 04 · COUNTING, SCREENING, JUDGING

ACTOR	WHAT IS CLASSIFIED	INSTRUMENT	USE	HUMAN STEP	OUTCOME
Pew Research Center	web pages (sample)	open-weight detector	MEASURE (AGGREGATE)	method and threshold published	population estimate; individual pages misclassified both ways
arXiv	submissions	moderation	SCREEN	staff and moderators	limit of two new submissions per month; one-year ban for in-controvertible unchecked AI content
NeurIPS 2026 Position Paper Track	submissions	commercial AI-text detector	SCREEN	decision routes varied by score band and stated conditions; additional considerations not listed for every group	178 desk-rejected (no appeal under standard circumstances); 123 offered an evidence route (version history)
Turnitin (vendor guidance)	student work	own AI-writing detector	GUIDANCE ON USE	human judgement required	result "should not be used as the sole basis for adverse actions"
University of Waterloo	student work	Turnitin AI detection	DISCONTINUED	—	feature switched off, Sep 2025
Curtin University	student work	Turnitin AI detection	DISCONTINUED	—	feature disabled from 1 Jan 2026
University of the Free State	student work	Turnitin AI detection	DISCONTINUED	—	discontinued from 1 Jul 2026
New York Supreme Court, Nassau County	one disciplinary determination	court review of the evidentiary record and process	SANCTION REVIEWED	court	determination annulled on case-specific evidentiary and procedural grounds; no general finding on detectors

The same kind of classifier is used to estimate prevalence, to screen submissions and, in some cases, to inform decisions about individuals.

NOTE Aggregate estimation and individual adjudication carry different error requirements (ICML 2024, Liang et al.).

SOURCES Pew Research Center, 20 Aug 2026; arXiv blog, 1 Oct 2026; NeurIPS blog, 2 Jun 2026; Turnitin guide; University of Waterloo; Curtin University; University of the Free State; *Matter of Newby v Adelphi Univ.*, 2026 NY Slip Op 26021

HOW TO READ

Each row is one actor and what it did with a classifier or with classifier evidence: measuring a population, screening submissions, issuing guidance, switching a feature off, or, in the court's case, reviewing the evidence and procedure behind one decision. The 'Use' column gives the purpose; the 'Human step' column shows where a person or body takes part.

References 1, 11, 7, 2, 3, 4, 5, 6, 8

2.2 Commercial reports and shared detection methods

Much of what is publicly known about AI-text prevalence comes from companies' own reports, several of them from companies that sell detection. This issue presents them as the companies' findings.

Graphite reports that 49.9% of sampled English-language articles in the first quarter of 2026 were primarily AI-written, and describes a plateau near half³². Originality.ai reports that 45.6% of its August 2026 sample crossed its threshold of 15% AI content³³. The thresholds differ, so the figures are not two readings of one quantity and are not compared here.

Pangram, whose open model Pew used, classified about 21% of the ICLR 2026 peer reviews it analysed as fully AI-generated. It reports that more than half showed some AI involvement²⁴, and that more AI involvement in a review went with a higher score²⁴. Pangram states a false-positive rate of 1 in 10,000, a figure that cannot be applied automatically to this corpus. It has publicly called on conferences to adopt AI detection²⁴. It both measures the phenomenon and sells a remedy. That does not make its figures wrong, but it is why they appear as "Pangram reports", and why reports built on the same detector are not counted as independent confirmation of one another^{23, 24}.

2.3 Universities step back from individual flags

Detector vendors' own guidance separates estimating from deciding. Turnitin's guide states that its AI-writing model "may not always be accurate". It suppresses numerical scores in the 1–19% range and requires at least 300 words of qualifying prose. Its result "should not be used as the sole basis for adverse actions against a student"².

Some universities have gone further and switched the feature off. The University of Waterloo discontinued Turnitin's AI detection in September 2025. It reported that its internal testing had flagged human-written text as 100% AI-generated, and gave its assessment that no product on the market effectively identified generative-AI writing³. Curtin University in Australia disabled the feature from 1 January 2026⁴. The University of the Free State in South Africa did so from 1 July 2026⁵. These are institutional decisions about one use, judging individual students. They are not findings that aggregate measurement is impossible^{8, 31}.

The same principle appears in other settings. At the start of the 2026–27 school year in the UAE, education expert Saeed Al Kaabi told *Emarat Al Youm* that detector results should be treated as indicators for further checking. He said they should be weighed alongside evidence such as drafts, sources, revision histories and the student's explanation of the work³⁴. An English teacher interviewed for the same report proposed following a piece of work from the first idea and draft, through sources and revisions, to a discussion of the final product with the student³⁴.

2.4 Bias in tested detectors

A 2023 Stanford study tested seven detectors on essays written by people. Across the seven detectors, the average false-positive rate on essays written by non-native English speakers for the TOEFL test was 61.22%, and 97% of those essays were flagged by at least one detector. Essays by US eighth-graders were classified almost perfectly³⁵. When the authors rewrote the TOEFL essays with more varied vocabulary, the disparity shrank in the tested sample³⁵. The finding concerns the detectors, essays and transformations tested in 2023. It does not describe every detector since, or all second-language writers. It shows how a detector's errors can fall unevenly across groups of writers.

2.5 Conferences: screening, then decision routes

In June 2026 the programme chairs of NeurIPS, a machine-learning conference, reported how they had handled AI-generated text in its Position Paper Track⁷.

In a preliminary analysis with the detector's default settings, 273 of 969 submissions (28.2%) received Pangram's maximum score⁷. For screening, the chairs then moved to shorter text windows, under which far fewer submissions scored at the top of the scale. The post does not explicitly map every decision group to a window configuration and score distribution, so we report each table in its own terms⁷. The post also gives 969 submissions in its text and 971 in a table⁷.

The chairs announced different decision routes:

- **178 submissions (18.4%)** were desk-rejected, in a decision "not subject to appeal under standard circumstances". For the highest-scoring group, the post lists no additional considerations⁷.
- **A further 123 (12.7%)**, in a lower score band, could submit version history showing a pre-AI draft, the AI-edited version and the final paper, for review by the track's team⁷.

The chairs explained their restrictions by reference to verification. Submitting AI-generated text for review, they wrote, "externalises the cost of verifying that work" onto reviewers⁷. This is a documented institutional policy. It is not an independent validation of the detector, and not a universal model for adjudication.

2.6 When a platform cannot tell

In late January 2026 a social network presented itself as a platform where AI agents post and humans watch. Owners registered an agent, which joined by following a set of instructions³⁶. The advertised participation model did not establish who operated each account.

The security firm Wiz reported that a key exposed in the site's client-side code, combined with missing database access controls, allowed unauthenticated read and write access to production data. The platform fixed the issue with Wiz's help³⁷. In the database as it stood before the first fix on 31 January 2026, Wiz counted about 1.5 million registered agents and about 17,000 human owners; it published the findings on 2 February (Figure 5)³⁷. Wiz also reported that the platform had no mechanism to verify whether a registered "agent" was an AI system or a person running a script³⁷. As Wiz's co-founder told Reuters: "You don't know which of them are AI agents, which of them are human"³⁸.

Registrations are not activity. A study by David Holtz of Columbia Business School observed the platform from 27 January to 31 January 2026, 11:30 UTC. It recorded 6,159 accounts with at least one observed post or comment in that window³⁹. The counts come from different dates and populations. Neither can be divided by the other to give a rate of activity or autonomy, and none establishes autonomous operation^{37, 39}. The episode shows the problem of this chapter in its plainest form: a system built for one kind of participant could not establish which kind it had.

— FIGURE 05 • 1.5 MILLION REGISTRATIONS, 17,000 OWNERS

1,500,000

registered agents

17,000

human owners

Database state before the first fix, 31 Jan 2026. Registrations and owner records — not activity.

About 1.5 million registered 'agents' and about 17,000 human owners, according to the Wiz investigation. Registrations do not measure activity or autonomy; the platform had no mechanism to verify whether an 'agent' was AI.

SOURCE Wiz, 2 Feb 2026.

HOW TO READ

Both numbers come from the same database, as it stood before the security flaw was fixed on 31 January 2026. They count registrations and owner records. Neither shows the total active population, or how many accounts were run by AI. A separate study recorded 6,159 posting or commenting accounts in its own early window.

References 37, 39

03

Marks that fade

Marking can identify covered outputs, but detection availability and robustness vary by implementation, access to keys, passage characteristics and transformations. Failure to detect a mark does not establish human authorship.

Opening: a mark answers a limited question

A provenance check looks for evidence associated with a particular marking system or signed record. A positive result can be evidence of a compatible marking process. On its own, it does not establish complete authorship or the truth of the content. A negative result does not establish human authorship^{9, 10, 40}.

So the practical questions are separate:

- which outputs are covered;
- what transformations the evidence can survive;
- who can perform the check;
- what a result establishes.

Figure 6 records the published arrangements of three providers in the same fields, without treating their different evaluations as a ranking.

3.1 Two mechanisms

Two mechanisms are discussed here. One is a signed provenance record attached to a file, under the C2PA "Content Credentials" standard. The other is a mark embedded in the content itself, invisible to people but readable by a detector. What either is worth as evidence depends on what is recorded, on the chain of trust behind it and on the verification method.

They fail differently. Provenance records can become separated from a file: provider documentation notes that metadata can be stripped in upload or download, or lost through a format change, resizing or a screenshot⁴¹. Stripping metadata does not by itself remove a mark embedded in the content; how well such a mark survives other transformations depends on the scheme (section 3.4). Where implemented, a compatible watermark or fingerprint can help recover the credentials associated with a file⁴². Section 3.4 shows that marks have weaknesses of their own.

3.2 Providers, one table

Figure 6 compares three providers of general-purpose models using the same defined fields: which media they mark, who can check, whether code or verification tooling is public, what robustness evaluations exist and what limits they state. Figures in it come from different implementations, corpora and operating points. The table does not rank the providers.

— FIGURE 06 · PROVIDERS, ONE TABLE

FIELD	ANTHROPIC ¹	GOOGLE ¹	OPENAI ¹
Image / video / audio mark	C2PA signatures on supported generated files; separate multimedia watermark not found	SynthID watermark (portal also checks partners' content)	SynthID watermark + C2PA metadata on images; supported audio since Jul 2026
Text mark	A version of SynthID-Text on supported models, worldwide subject to model/feature/platform support	SynthID-Text	Text watermark: ChatGPT/Codex text marked by default in the EU only; API opt-in, off by default
Who can check (hosted)	Files: free content checker + C2PA tools; text: detection API in private preview	Public portal for images, video, audio	Images and audio: public verifier + API; text detector: approved researchers only
Detector code / verification tooling	Text production-detector source: not found. File credentials: verification with compatible C2PA tooling is documented, including the open-source c2patool.	SynthID-Text reference code public, incl. detection ("not intended for production")	Announced intent to open-source
Published text robustness evaluation	No Claude-specific numerical robustness evaluation comparable in reporting form to the other table entries was found	Peer-reviewed (Nature 2024): evaluation incl. edits and paraphrasing; numbers in Supplementary C.6	Provider's own (5 Oct 2026), 1% FPR target: ~80% detected at 200 tokens, ~95% at 400 tokens (psychology); lower for mathematics. Synonym substitution on 400-token English answers: 10% of words → 92% to 66%; 25% → 17%

FIELD	ANTHROPIC ¹	GOOGLE ¹	OPENAI ¹
Self-stated limits	Short text weaker; mark is not complete authorship; missing credentials not human origin	Watermarks weakened by edits; vulnerable to stealing, spoofing and scrubbing; coordination among providers can broaden coverage	Absence of a mark does not establish human authorship
FOOTNOTE 1 Each provider's column had an independent reviewer; see Method. FOOTNOTE 2 Figures come from different implementations, corpora and operating points. No ranking is implied. FOOTNOTE 3 "Not found" means not located in the stated review; it does not establish non-existence, concealment, unlawfulness or non-compliance. SOURCES Anthropic: Claude Help Centre, "How Claude marks AI-generated content"; Claude Platform documentation (code execution tool); Anthropic, text-watermark announcement — retrieved Oct 2026 Google: Google blog, 7 Oct 2026 (SynthID Detector); Dathathri et al., Nature 634 (2024); github.com/google-deepmind/synthid-text OpenAI: OpenAI, "Advancing content provenance" (19 May 2026, update 31 Jul 2026); OpenAI, "Our approach to EU text provenance rules" (5 Oct 2026)			
HOW TO READ Each column is one provider, and each row asks all three the same question. The cells summarise what the provider or its researchers have published. The test results in the robustness row come from different methods and data, so they cannot be compared across columns. The table does not rank the providers. References 9, 10, 40, 41			

3.3 Text under edits

Text and image marking face different constraints, and their robustness depends on method, content, transformation and operating point. The evaluations reviewed here do not establish a general ranking between media.

For text, the selected published evaluations, one a provider's own report and one a peer-reviewed paper, report that marks weaken as text is edited or paraphrased. Their settings differ and their numbers are not comparable; both are in Figure 6^{9,10}. The peer-reviewed paper identifies further limitations: marks can be stolen, spoofed or scrubbed. A detector for one marking scheme does not establish the origin of output from tools outside that scheme. Coordination between providers can broaden coverage¹⁰.

3.4 Image marks under attack

Academic work shows how image marks can be removed or forged. A 2023 study found a trade-off between evading a watermark and spoofing one. Purification with a diffusion model removed low-perturbation watermarks, and a model-substitution attack removed high-perturbation ones⁴³. A paper at CVPR 2026 showed a "visual paraphrase" attack that strips both visible and invisible marks while keeping what the image shows⁴⁴.

Mechanisms can also interfere with each other. An October 2026 preprint by authors at a provenance-technology company reports that, in its tested configuration, embedding a watermark can move an image's ISCC content fingerprint past the threshold at which it still matches the unmarked original. That is fingerprint drift, not a failure to detect the watermark. The authors examine whether a registry should hold the fingerprint of the original or of the marked asset⁴⁵. They also report 53 entries in the C2PA soft-binding registry snapshot they examined, dated 16 September 2026: 44 watermarks and 9 fingerprints. Inclusion in that list does not itself establish robustness or certification⁴⁵.

3.5 Who can check

Access to checking takes several forms:

- hosted services;
- public programming interfaces;
- restricted previews;
- research-only access;
- published reference implementations.

None of them automatically supplies the keys or permissions needed to verify every production output. Figure 6 records the arrangements of the implementations reviewed^{9,10,41}.

3.6 What labels communicate, and what they do not

A label communicates an asserted production history. Whether viewers understand it, whether they believe the content shown with it, and whether they intend to share that content are separate outcomes.

The evidence runs in both directions. A January 2026 briefing by the UK House of Commons Library summarised experiments in which labelling headlines as AI-generated "lowered their perceived accuracy", including for true headlines; it also noted that findings vary⁴⁶. A 2025 study in *PNAS Nexus*, with fieldwork in late 2023, ran two preregistered experiments with 7,579 US participants shown misleading AI-generated images. "All of the labels we tested significantly decreased participants' belief in the presented claims." Effects on stated engagement intentions were weaker and varied by label design⁴⁷.

Australia's voluntary government guidance for business discusses the different properties of labels, watermarks and metadata, and their possible combined use⁴⁸. A label is a statement about provenance. It is not a test of truth, and its effects on belief need to be measured, not assumed^{46,47}.

04

Who must mark what

The specific California latent-disclosure provision and India's SGI-specific framework focus on audiovisual content. EU and Korean provisions also address text, but their duties, covered actors, exceptions and commencement arrangements differ. This is a comparison of specified obligations, not a ranking of jurisdictions' overall AI regulation.

Opening: one date, two regimes

On 2 August 2026 two transparency regimes became applicable on the same day. In the European Union, Article 50 of the AI Act began to apply⁴⁹. In California, the AI Transparency Act became operative on the date to which the legislature had deferred it⁵⁰. The two regimes differ in covered actors, media, obligations, exceptions and operative dates. Figure 7 sets out the instruments we located with those fields side by side.

4.1 European Union

Article 50 imposes provider-marking and deployer-disclosure duties within defined scopes⁴⁹:

- **Providers** must ensure that the output of generative systems is marked in a machine-readable format and detectable as artificially generated. The technical solutions must meet specified effectiveness, interoperability, robustness and reliability requirements as far as technically feasible. The exceptions include an assistive function for standard editing, and systems that do not substantially alter the input data.
- **Deployers** must disclose deepfakes, with adapted arrangements for evidently artistic, creative or satirical works.
- **Deployers** must also disclose AI-generated text published to inform the public on matters of public interest, unless it has undergone human review or editorial control and a person holds editorial responsibility for it. That exception concerns the deployer's disclosure; it does not exempt the provider from marking.

The European Commission published a final Code of Practice on marking and labelling on 10 June 2026. It is a voluntary means of showing compliance, not a regulation⁵¹. On 20 July 2026 the Commission also published guidelines on the scope of the Article 50 transparency obligations¹⁰⁵. In July the Digital Omnibus, Regulation (EU) 2026/1744, gave systems placed on the market before 2 August 2026 until 2 December 2026 to comply with the marking duty⁵². Article 99 sets maximum penalties, which are not automatic. For breaches of Article 50 the maximum is €15 million or 3% of worldwide turnover, whichever is higher, with the lower of the two applying to specified smaller enterprises⁴⁹.

4.2 United States: California

California regulates by statute. The AI Transparency Act requires covered providers to embed a latent disclosure in the image, video and audio content they generate^{50, 53}. On 30 September 2026 SB 1000 was signed as an urgency statute (Chapter 861). It removed the user-count threshold that had limited the law to large providers, and it requires providers to offer a "disclosure verification tool" at no cost to the user⁵³. Specified duties for covered platforms begin in 2027, subject to the statute's feasibility and compatibility provisions. Covered newly produced capture devices are addressed separately from 2028^{50, 54}.

The latent-disclosure provision lists image, video and audio. This provision does not cover text, which is not the same as saying California leaves AI text unregulated⁵³.

4.3 India and South Korea

India's amendment to its IT Rules was notified on 10 February 2026 and took effect on 20 February. It introduces labelling and provenance duties for "synthetically generated information" (SGI) within a defined scope. The much-reported three-hour window applies to specified court orders and authorised government notices; it is not a universal takedown deadline⁵⁵. Under the government's FAQ, text-only AI output is not SGI, although general obligations on unlawful information still apply to text⁵⁵.

South Korea's AI Basic Act commenced on 22 January 2026. Fines of up to 30 million won are tied to enumerated infringements, such as failures of notification or of compliance with corrective orders, not to each missing mark. The government has announced a separate enforcement grace period⁵⁶. Guidance from the Ministry of Science and ICT sets different obligations and acceptable disclosure methods according to the type and presentation of the output, with stronger requirements for realistic synthetic content⁵⁶.

4.4 China

China's labelling measures and later Cyberspace Administration programmes appear in Figure 7 as located but not cleared, from their official Chinese text⁵⁷.

4.5 United Kingdom, Australia

In January 2026 the House of Commons Library stated: "There is no UK legislation that requires AI-generated content to be labelled as such."⁴⁶ That is the briefing's account at its date of publication.

Australia has two separate instruments. The National AI Centre's guidance "Being clear about AI-generated content" is voluntary. It discusses labelling, watermarking and metadata⁴⁸. Separately, an industry standard under the eSafety Commissioner requires certain high-impact generative-AI services to have systems that "differentiate AI outputs generated by the model"⁵⁸.

Japan and Brazil appear in Figure 7 as located but not cleared⁵⁷.

4.6 Gulf states reviewed

We reviewed five Gulf states: the UAE, Saudi Arabia, Qatar, Bahrain and Oman. The measures we located address AI-generated content through media regulation, content standards and published AI policies. Saudi Arabia's published guidance was located but not cleared for detailed description (Figure 7)⁵⁷.

In the UAE, WAM reported on 25 September 2025 that the UAE Media Council had stated that using AI to depict national symbols or public figures without prior official approval violates media content standards, and that misuse of AI to spread misinformation, incite hatred or defame others is subject to the media violations regulation⁵⁹. The Council is named here as the issuer of that statement. The National Media Authority, established under Federal Decree-Law No. 11 of 2025 and announced in December 2025, has since replaced it¹⁰³. Khaleej Times and Gulf News reported at the end of September 2026 that the Authority had restated, in a post on its official account, that creating images or videos of national symbols or public figures with AI without prior official approval breaches media content standards; we did not retrieve the post itself¹⁰⁴. In August 2026 WAM reported that the International Centre for Combating Media Disinformation (ICCMD) had announced the development of an AI-supported system for monitoring, analysing and verifying digital content, together with an International Charter for Media Integrity and Combating Misinformation⁶⁰.

Qatar's Ministry of Communications and Information Technology has published AI ethical guidelines, which state that they are voluntary and not legally binding⁶¹. Bahrain's Information & eGovernment Authority issued a General Policy for the Use of AI in July 2025⁶². Oman's Ministry of Transport, Communications and Information Technology published a General Policy for the Safe and Ethical Use of AI Systems on 1 April 2025⁶³. Chapter 7 states the bounds of our search in the region.

4.7 Elsewhere

Outside the jurisdictions above, the instruments we were able to verify take several forms:

- **Statute.** Vietnam's Law on Artificial Intelligence, in force since 1 March 2026, separates provider marking from employer disclosure for specified generated or edited outputs. Its conditions are set out in Article 11 and its implementing arrangements⁶⁴.

- **Election rule.** Singapore prohibits publishing a "realistic but false audio, visual or audiovisual representation" of a candidate during the election period. A label does not make such material permissible⁶⁵.
- **Recommendations and guidance.**
- An Indonesian communications-ministry official has publicly recommended labels or disclaimers on AI-generated content⁶⁶.
- Malaysia's national AI governance guidelines discuss disclosure among suggested consumer-protection measures⁶⁷.
- Kenya's Media Council guide asks media enterprises to inform audiences of their AI use⁶⁸.
- Ukraine's media guidance recommends "clear separation between human-made and AI-generated content"⁶⁹.
- Nigeria's communications-sector incident-response team has issued an advisory on verification procedures against AI-driven impersonation and deepfakes⁷⁰.

Figure 7 also lists, as located but not cleared, the instruments below. The reason for each is given in the figure:

- **Legislation:** Russia and Kazakhstan⁵⁷.
- **Guidance and recommendations:** Mexico, Argentina and Egypt⁵⁷.
- **Programme conditions:** Colombia⁵⁷.

— FIGURE 07 • WHO MUST MARK WHAT

JURISDICTION	INSTRUMENT	ISSUER	COVERED ACTOR	MEDIUM	DUTY	STATUS	RELEVANT DATE (TYPE STATED)	SOURCE
STATUTE / REGULATION								
China	Labelling Measures and later CAC programmes	Cyberspace Administration of China	—	—	—	official text located (Chinese) LOCATED, NOT CLEARED	—	cac.gov.cn
European Union	AI Act Art. 50 (as amended by Reg. 2026/1744)	EU	providers; deployers	generative output; deepfakes; public-interest text	providers: machine-readable marking (exceptions incl. assistive standard editing). Deployers: disclose deepfakes (adapted for evidently artistic, creative or satirical work); disclose AI-generated text published to inform the public on matters of public interest, unless it has undergone human review or editorial control AND a person holds editorial responsibility. That exception applies only to the text disclosure, not to deepfakes or provider marking	applies	applicable from 2 Aug 2026; legacy systems by 2 Dec 2026	EUR-Lex; AI Act Service Desk
India	IT Rules amendment 2026	MeitY	intermediaries; SGI tools	synthetically generated information (text-only excluded)	label and provenance for SGI	in force	in force 20 Feb 2026	MeitY FAQ
Kazakhstan	AI Law No. 230-VIII	Parliament	—	—	—	record located; full text not retrieved LOCATED, NOT CLEARED	—	adilet.zan.kz
Russia	Federal Law 243-FZ	Federal Assembly	—	—	—	record located; operative text not verified LOCATED, NOT CLEARED	—	publication.pravo.gov.ru
South Korea	AI Basic Act + MSIT guidance	National Assembly; MSIT	providers	outputs; realistic synthetic content	marking and notice, differentiated by output type	commenced; grace period announced	commenced 22 Jan 2026	MSIT
United States — California	AI Transparency Act as amended (SB 1000; AB 853; AB 2713)	State of California	providers (no user threshold); covered platforms from 2027	image, video, audio (latent disclosure)	latent disclosure; free disclosure verification tool	operative	operative 2 Aug 2026; platforms from 1 Jan 2027; capture devices from 2028	California legislature (chapters texts)
Vietnam	Law on AI No. 134/2025/QH15	National Assembly	providers; deployers	specified generated or edited outputs	providers: machine-readable marking; deployers: disclosure (Art. 11 conditions)	in force	in force 1 Mar 2026	Government of Vietnam (English text)

JURISDICTION	INSTRUMENT	ISSUER	COVERED ACTOR	MEDIUM	DUTY	STATUS	RELEVANT DATE (TYPE STATED)	SOURCE
Brazil	PL 2338/2023 (general AI bill)	Chamber of Deputies	—	—	—	pending in Chamber LOCATED, NOT CLEARED	—	camara.leg.br
ELECTION RULE								
Brazil	TSE Resolution 23.755/2026	Superior Electoral Court	—	—	—	official text located (Portuguese) LOCATED, NOT CLEARED	—	tse.jus.br
Japan	Revised Public Offices Election Act (Law No. 58)	Diet	—	—	—	official record located (Japanese) LOCATED, NOT CLEARED	—	sangiin.go.jp
Singapore	Prohibition of realistic but false candidate depictions	Elections Department	publishers incl. redistribution	audio, visual, audiovisual	prohibited during election period (a label does not exempt)	in force	applies during the election period	Elections Department Singapore
SECTORAL / MEDIA STANDARD								
Australia	eSafety DIS Standard	eSafety Commissioner	high-impact generative-AI services	AI outputs	differentiate AI outputs	in force	—	eSafety Commissioner
Colombia	Territorios al Aire 2026 programme conditions	MinTIC	—	—	—	official text located (Spanish) LOCATED, NOT CLEARED	—	mintic.gov.co
UAE	Media content standards (Council statement, Sep 2025)	UAE Media Council (2025); National Media Authority, its successor (restatement, Sep 2026, as reported)	media and content creators	AI depictions of national symbols and public figures	prior official approval for AI depictions of national symbols and public figures; AI misuse to spread misinformation, incite hatred or defame others subject to the media violations regulation	statement reported by WAM; restatement reported by press	statement 25 Sep 2025; restatement late Sep 2026	WAM; Khaleej Times; Gulf News
GUIDANCE / POLICY								
Australia	Being clear about AI-generated content	National AI Centre	businesses	—	discusses labelling, watermarking, metadata	voluntary	published Nov 2025	industry.gov.au
Bahrain	General Policy for the Use of AI	Information & eGovernment Authority	—	—	general AI policy; marking clauses not verified	issued CLEARED (ISSUANCE)	issued Jul 2025	iga.gov.bh
Egypt	National Guidelines for Generative AI	Government of Egypt (ai.gov.eg)	—	—	—	located; full text not retrieved LOCATED, NOT CLEARED	published Mar 2026	ai.gov.eg

JURISDICTION	INSTRUMENT	ISSUER	COVERED ACTOR	MEDIUM	DUTY	STATUS	RELEVANT DATE (TYPE STATED)	SOURCE
Indonesia	Public recommendation by a ministry official	Communications ministry	content creators	AI-generated content	label or disclaimer (recommended)	recommendation	statement Jun 2025	ANTARA
Kenya	Media guide on the use of AI	Media Council of Kenya	media enterprises	—	inform users of AI use	sectoral guidance	—	Media Council of Kenya
Malaysia	National Guidelines on AI Governance & Ethics	MOSTI	—	—	disclosure among suggested measures	guidance	—	MOSTI
Mexico	Recommendations for election coverage	INE	—	—	—	official text located (Spanish); non-binding LOCATED, NOT CLEARED	issued Aug 2026	ine.mx
Oman	General Policy for the Safe and Ethical Use of AI Systems	MTCIT	—	—	general AI policy; marking clauses not verified	published CLEARED (PUBLICATION)	published 1 Apr 2025	mtcit.gov.om
Qatar	AI ethical guidelines	MCIT	—	—	—	voluntary, non-binding (self-stated)	—	MCIT
Saudi Arabia	Generative AI public guidance; deepfake guidance	SDAIA	—	—	—	located; full text not retrieved LOCATED, NOT CLEARED	—	sdaia.gov.sa
Ukraine	Guidelines for responsible AI use in media	National Council (with ministries)	media	—	clear separation of human-made and AI-generated content (recommended)	guidance	—	National Council of Ukraine
Instruments we located and verified. A comparison of specified obligations, not a ranking of countries' overall AI regulation. Grey rows: official texts located but not cleared for detailed description; only metadata is shown.								
NOTE United Kingdom: a January 2026 House of Commons Library briefing stated that no UK legislation requires AI-generated content to be labelled.								
SOURCES Official instruments and issuer publications named in the Source column.								
HOW TO READ Rows are grouped by type of instrument: statutes and regulations, election rules, sector standards, and guidance or policy. Read along a row to see who must do what, for which kind of content, and from when. The date column states what kind of date it is: application, entry into force, publication or issuance. Grey rows are official texts we located but could not verify in enough detail to describe; only their metadata is shown. The table compares specified obligations; it does not rank countries. References 49, 50, 53, 54, 55, 56, 57, 58, 59, 103, 104, 61, 62, 63, 64, 65, 66, 67, 68, 69, 46, 48								

05

Who checks the volume

Unchecked submissions can shift verification work to moderators and reviewers. The net effect of AI assistance depends on the workflow and has not been established by the cross-sector examples reviewed here.

The core of this chapter is limited to science (sections 5.1–5.2). Sections 5.3–5.5 add separate descriptive evidence and conditional mechanisms from other sectors; they do not show that the same cost transfer operates across them.

Opening: nine thousand tickets

arXiv's announcement of its rate limit set out the workload behind it. arXiv reported that September 2026 submissions generated nearly 9,000 support tickets for its staff and moderators¹¹. It called the limit a stopgap until new submission practices and moderation tools are in place¹¹.

Unchecked submissions can impose additional checking work on the people who receive them. The NeurIPS chairs described the same mechanism in their own terms (section 2.5)⁷.

5.1 Moderators

The measures are not a blanket prohibition on AI assistance:

- Since October 2025, review and position papers in arXiv's computer-science section have needed prior peer review. In January 2026 arXiv changed how new authors qualify for endorsement⁷¹.
- The chair of its editorial advisory council has said on record that authors who submit incontrovertible, unchecked AI-generated content face a one-year ban. The rule targets unchecked content, not the use of AI⁷².
- Conferences set their own per-author submission caps, and these differ by venue, edition and track⁷³. Not all were set in response to AI, and none of these measures has a published effectiveness evaluation in our evidence.

Where the load comes from matters¹¹. If a small minority of authors generates a large share of it, the problem is a heavy tail. That would not be evidence of a general decline in the quality of science.

5.2 Reviewers

Peer reviewers sit on both sides of the problem: they receive AI-assisted papers, and some use AI to review them. Pangram reports the following²⁴:

- it classified about 21% of the ICLR 2026 reviews it analysed as fully AI-generated;
- reviews it classified as AI-generated were longer and had lower information density by its measures;
- more AI involvement went with higher scores.

AI assistance in reviewing is not only a cost. In a randomised study at ICLR 2025, reviewers received AI-generated feedback on their draft reviews. In blinded assessment of the evaluated revisions, the revised reviews were preferred in 89% of cases⁷⁴. That result concerns those revisions. It is not an estimate for all AI-written reviews, or for total reviewing workload.

IN SHORT *at selected machine-learning venues, detector-based manuscript screening coexists with estimated or documented LLM assistance in reviewing. These observations do not establish an autonomous paper-review-decision cycle.*

5.3 Recruiters, publishers and reviews

Other sectors supply separate descriptive evidence. It is measured in different ways, and it is not a measurement of the mechanism described in sections 5.1–5.2.

- **Recruitment.** Robert Half reports, from a June 2026 survey of 500 hiring managers and 1,000 professionals in the UK, that 67% of hiring managers said the need to screen for AI-generated materials had increased the time it takes to hire. It reports 27% for a category it calls "distinguishing between AI-generated and authentic CVs"; that wording should not be read as establishing that AI assistance makes a CV inauthentic⁷⁵. These are reported experiences, not measured effects. In a separate US release the firm notes that "not all AI-generated applications are inaccurate or misleading"⁷⁶. Robert Walters reports that 42% of employers in its Middle East survey were receiving more irrelevant applications. That is a regional company survey, not a representative estimate for any one country⁷⁷.
- **Publishing.** Amazon's self-publishing service requires authors to disclose AI-generated text, images and translations to Amazon, though not AI-assisted content. Its help page, retrieved in October 2026, limits submissions to two titles per book format each week⁷⁸. The disclosure goes to the platform; it is not a label for readers.
- **Reviews.** The US Federal Trade Commission's rule on fake reviews and testimonials took effect on 21 October 2024. It covers fictitious reviews, including AI-generated ones, but does not prohibit every review written with AI help⁷⁹. The FTC has since brought at least one complaint under it⁷⁹. In December 2025 it set aside an earlier order against an AI writing service, which is therefore no longer in force⁸⁰.
- **Pangram's review sample.** In a May 2026 sample of 30,000 front-page reviews of 500 bestselling Amazon products across ten categories, Pangram classified 909, about 3%, as AI-generated with high confidence⁸¹.

How a review was written, whether the experience it describes is genuine, and whether it breaks the law are three separate questions.

IN SHORT *across documented news, recruitment, publishing and review contexts, organisations report AI-associated content or have introduced disclosure, authenticity and verification controls. The evidence uses different populations and methods; it does not establish a common synthetic-content share, a uniform transfer of verification costs, or the net effect of AI assistance across sectors.*

5.4 Audiences that are partly automated

Advertising systems use traffic and engagement measures whose interpretation can be affected by automation.

Vendors who measure automated traffic report its rise:

- **Thales/Imperva** reports that automated traffic made up 53% of the web traffic it observed in 2025, up from 51% in 2024, with "bad bots" at 40%⁸². It now counts AI agents as a third category of automated traffic⁸². It says this share makes engagement, conversion and demand signals harder to interpret⁸².
- **HUMAN Security** reports that automated traffic across its customers grew eight times faster than human traffic, and AI-driven traffic by 187% over 2025⁸³. HUMAN frames the question as whether traffic can be trusted, not whether it is human⁸³.

Published estimates of what this costs advertisers cannot be treated as estimates of one common quantity without harmonisation:

- **Spider Labs** reports a 5.58% fraud rate in its sample of 4.006 billion paid clicks in the first half of 2026. Its \$25.3 billion figure extrapolates that rate to global spend for the half-year⁸⁴.
- **Fraudlogix** reports that its signals flagged 20.64% of impressions as invalid in 2025. That counts impressions, not clicks, and a flag is not a loss⁸⁵.
- **DoubleVerify** reports 140% more connected-TV fraud schemes and variants in the first quarter of 2026 than in the first quarter of 2025. That counts schemes, not impressions or dollars⁸⁶.

Figure 8 sets out what each figure measures.

— FIGURE 08 • DIFFERENT SAMPLES, PERIODS AND UNITS							
VENDOR	REPORTED FIGURE	MEASURE TYPE	PERIOD	UNIT	SAMPLE	EXTRAPOLA-TION	SOURCE
Spider Labs	5.58% fraud rate; \$25.3bn	rate (share of paid clicks) + extrapolated amount	H1 2026	paid clicks	4.006bn paid clicks	global half-year extrapolation	Spider Labs release, 2026
Fraudlogix	20.64% flagged invalid	rate (share of impressions flagged)	calendar 2025	impressions	105.7bn impressions	none stated	Fraudlogix, Ad Fraud Statistics 2026
DoubleVerify	+140% CTV fraud schemes and variants	growth rate of a count (schemes and variants)	Q1 2026 vs Q1 2025	identified schemes	vendor detections	none	DoubleVerify, 7 May 2026
Thales/Imperva	53% automated traffic; 40% bad bots	rate (share of observed traffic)	2025	web traffic	traffic observed by the vendor	none	Imperva Bad Bot Report 2026
HUMAN	automated traffic grew 8x faster than human traffic	comparison of growth rates	2025	traffic growth rate	customer traffic	none	HUMAN, 2026 State of AI Traffic
Vendors measure different things (rates of clicks, impressions or traffic; growth in a count of schemes; relative growth) over different periods and samples. The figures are not a range for one quantity.							
SOURCES Vendor publications named in the Source column; figures are vendor-reported.							

HOW TO READ

Each row is one company's headline figure. Read the measure type, unit, period and sample before the figure itself. Some figures are rates (a share of clicks, impressions or traffic), one is the growth of a count of schemes, and one compares growth rates. Rates, counts and growth rates measure different things, so they cannot be added, averaged or read as a range.

References 82, 83, 84, 85, 86

IN SHORT *published vendor figures are not directly comparable because their observation periods, sampling frames, traffic units and loss-extrapolation methods differ. The evidence reviewed here does not establish a like-for-like range for global ad-fraud losses. Automation is not the same as fraud either. The vendors' own categories separate good bots, bad bots and AI agents⁸².*

5.5 Learning from contaminated signals

Ad systems also learn from clicks, not only pay for them. Which ads are shown, and to whom, depends on estimated click-through and conversion rates. A 2005 paper by Nicole Immorlica, Kamal Jain, Mohammad Mahdian and Kunal Talwar showed that common methods of learning click-through rates from observed clicks are vulnerable to fraudulent clicks. It also identified estimators that resist such fraud under its assumptions⁸⁷. A 2026 paper at SIGIR found that noisy conversion labels degrade conversion-rate prediction, a related but different task⁸⁸.

These are component mechanisms. They show that automated engagement *could* distort the optimisation of advertising systems if it is counted and learned from, and that the design of the estimator matters. The cited studies do not establish the complete chain from bot conversions, through learning, to the allocation of more fraudulent traffic.

06

Assisted scores, unaided skills

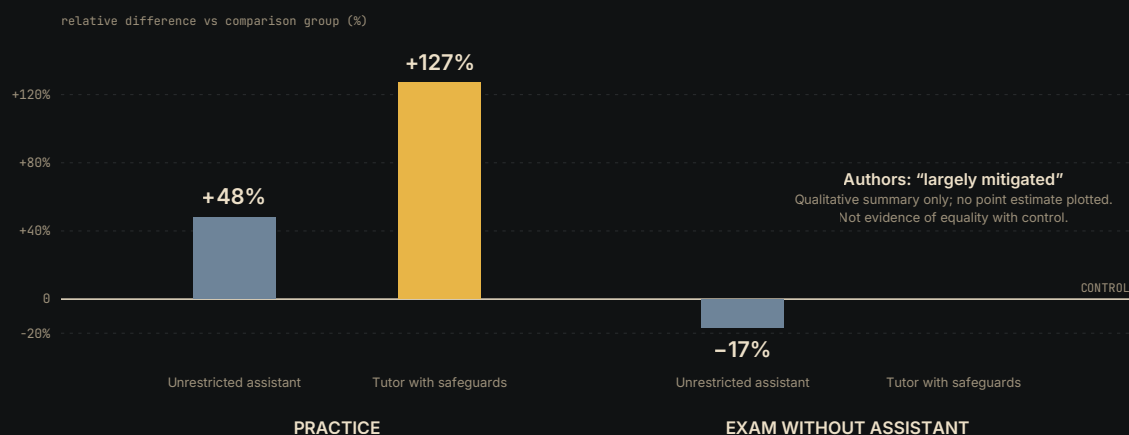
In the learning studies reviewed, AI assistance improved scores on assisted tasks. The studies did not establish a corresponding advantage on their assessed unassisted outcomes; the outcomes and interventions differed. In the mathematics study, unrestricted access was followed by lower unassisted exam scores, which the authors report a safe-guarded tutor largely mitigated. No generalisation beyond learning.

Opening: two scores for the same student

In a field experiment with nearly 1,000 high-school mathematics students in Turkey, researchers gave one group unrestricted access to a general-purpose AI chat assistant during practice sessions^{12,13}. A comparison group did not receive the assistant during those sessions. In relative terms, the first group's practice scores were 48% above the comparison group's. Then the assistant was removed for the exam, and the same students scored 17% *below* the comparison group.

A second group used a version of the assistant designed to give pedagogical support, through questions and hints, rather than simply to supply answers. Its practice scores were 127% above the comparison group's. On the unassisted exam, the authors describe the negative effect as "largely mitigated". That is not a demonstrated equivalence with the comparison group (Figure 9)¹².

FIGURE 09 • PRACTICE VERSUS EXAM



Unrestricted access raised practice scores by 48% and was followed by exam scores 17% below the control group once access was removed. For a tutor designed not to hand out answers, the authors report the loss was largely mitigated.

A separate randomised study (BJET, 2025; 117 students): the AI-assisted group showed the greatest improvement in essay scores during revision; the study did not detect significant between-group differences in knowledge gain or transfer.

SOURCE Bastani et al., PNAS 122(26), 2025; Fan et al., British Journal of Educational Technology 56(2), 2025.

HOW TO READ

The bars show how much higher or lower each group scored than a comparison group without AI. The left pair covers practice sessions with AI; the right pair covers the exam taken without it. For the hint-based tutor, the exam result is given only as the authors describe it; no point estimate is plotted, and the label is not evidence of equality with the comparison group.

References 12, 14

6.1 Two controlled studies

A separate randomised study of 117 university students, published in the *British Journal of Educational Technology*, compared four kinds of support for a writing task: an AI chatbot, a human expert, writing-analytics tools, and none. The AI-assisted group showed the greatest improvement in essay scores during revision. The study did not detect significant between-group differences in knowledge gain or transfer¹⁴.

The two studies measure different outcomes, and their results are not pooled here. In both, the assistance improved scores on the assisted task. Neither established a corresponding advantage on unassisted outcomes. In the mathematics study the safeguarded intervention had a different outcome from unrestricted assistance; the study does not isolate the causal contribution of each design feature¹². These are learning studies, and they say nothing directly about other kinds of work.

6.2 The wider evidence

The OECD's 2026 *Digital Education Outlook* distinguishes task-performance gains from learning gains. It associates better learning outcomes with uses designed with a pedagogical purpose⁸⁹.

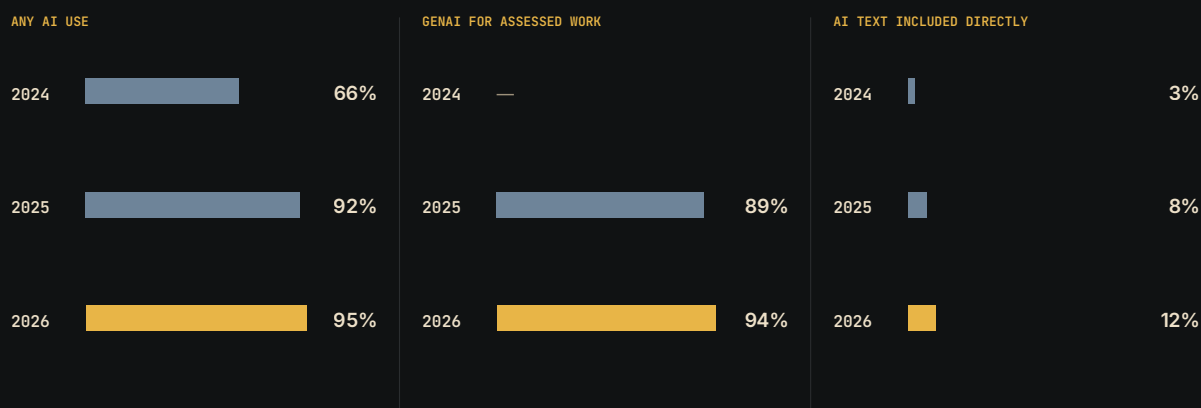
6.3 What students report

Use is widespread. In a Pew Research Center survey of 1,458 US teenagers published in February 2026, 54% said they use AI chatbots for schoolwork, and 10% said they use them for all or most of it⁹⁰. This issue does not use the survey to build a multi-year series.

HEPI's 2026 report draws on a December 2025 survey of 1,054 full-time UK undergraduates (Figure 10)⁹¹:

- 95% use AI in some form, against 66% in 2024 and 92% in 2025;
- 94% use generative AI in assessed work, against 89% a year earlier;
- the share who said they had put AI-generated text directly into assessed work was much lower but rising: 3% in 2024, 8% in 2025, 12% in 2026⁹¹.

— FIGURE 10 • UK UNDERGRADUATES AND AI, 2024-2026



95% of surveyed UK undergraduates used AI in some form; 12% said they included AI-generated text directly in assessed work (2026).

SOURCE HEPI, Student Generative AI Survey 2026 (survey of 1,054 full-time UK undergraduates, December 2025).

HOW TO READ

Each panel is one survey question, asked in 2024, 2025 and 2026 of full-time UK undergraduates. The first shows any use of AI, the second its use in assessed work, the third including AI-generated text directly in assessed work. No 2024 value is shown for assessed work.

References 91

Students disagree about the effect. 49% described it as positive, 16% as negative and 35% as making no difference. One student among those reporting a negative effect wrote: "I'm not using my brain at all"⁹¹.

6.4 Knowledge workers

A 2025 study surveyed 319 knowledge workers about their own use of AI tools. Workers who expressed more confidence in AI reported engaging in less critical thinking. Those more confident in their own abilities reported more⁹². The study measures what workers report, not what they do.

6.5 Earlier evidence, flagged

Two often-cited sources need flags. A 2025 MIT Media Lab study recorded brain activity while participants wrote essays with an AI model, with a search engine, or unaided. Participants who had used the AI model struggled to quote accurately from their own essays. The study is a preprint and has not been peer-reviewed. Its fourth-session findings concern 18 participants, of 54 in the first three sessions, and should be read within that limited design⁹³. Separately, the OECD's 2024 survey of adult skills found that literacy and numeracy had declined or stagnated in most participating countries⁹⁴. These trends do not identify an effect caused by generative-AI use⁹⁴.

6.6 UAE education: published measures and stated safeguards

Official announcements describe the introduction of AI education in UAE public schools from the 2025–26 academic year, and responsible-use measures at federal and emirate level. This section reports the measures described in the cited announcements; it does not evaluate implementation or educational outcomes.

- **Ministry of Education.** The Ministry describes AI as taught in public schools from kindergarten to Grade 12 within the existing Computing, Creative Design and Innovation subject⁹⁵.
- **Abu Dhabi.** The Department of Education and Knowledge (ADEK) describes guardrails "designed to prevent over-reliance on AI and protect students' cognitive agency" across more than 170 private schools, alongside a baseline assessment to track progress⁹⁶.
- **Sharjah.** WAM reported on 14 October 2026 that the Sharjah Private Education Authority (SPEA) had launched an ethical framework. The announcement describes requirements concerning human oversight, preservation of students' original work, disclosure of AI use where appropriate, and verification of AI-generated information and its sources⁹⁷.
- **Ras Al Khaimah.** In an August 2025 statement reported by *Khaleej Times*, the Department of Knowledge identified teacher training in responsible use as its first priority⁹⁸.
- **Dubai.** In statements reported on 1 October 2026, KHDA said it was preparing a framework centred on when AI strengthens learning, and that AI use should be "developmentally appropriate and supports meaningful learning"⁹⁹.

Gulf News (February 2026) and Education Middle East report that the Ministry of Education's 2026 classroom guide requires teacher-supervised use, restricts generative AI for pupils under 13 or below Year 7, as termed in those reports, and excludes it from formal examinations. We did not locate the guide itself¹⁰⁰.

School-level accounts describe approaches to responsible use. An English teacher interviewed by *Emarat Al Youm* proposed following a piece of work from the first idea and draft, through sources and revisions, to a discussion of the final product³⁴. *Rest of World* described an example activity at a Dubai school in which a teacher might show children something an AI system got wrong and ask them to identify what is incorrect or unfair¹⁰¹. These accounts are not evaluations of implementation across schools.

These are announced measures and stated objectives. They are not a tested replication of the studies in section 6.1.

What is not established

This chapter records what the evidence does not support, so that the earlier chapters are read at their actual strength.

7.1 Claims we withdrew

Two working hypotheses did not survive verification:

- **"Performance without the person"**. The slogan had no operationally testable meaning. The learning evidence is reported in Chapter 6 without it.
- **"Identity became the scarce asset"**. The record did not measure scarcity, price or substitution of identity. A narrower observation remains: some workflows require verifiable provenance, author responsibility or authorised agency, and those requirements should be measured separately.

7.2 Claims we narrowed

The left-hand column records withdrawn working claims. They are not findings and should not be quoted as such.

TOPIC	WITHDRAWN CLAIM (NOT A FINDING)	WORDING IN FORCE
Detection and decisions (Ch. 2)	The cited measurement and decision settings all rested on contested vendor classifiers	Authorship and traffic measurements are method- and population-dependent. Validated aggregate estimates can be informative even when an individual classification is insufficient for disciplinary or other adverse action
Marking (Ch. 3)	Detection access is restricted by every provider	Marking can identify covered outputs, but detection availability and robustness vary by implementation, access to keys, passage characteristics and transformations. Failure to detect a mark does not establish human authorship
Checking costs (Ch. 5)	Verification cost is pushed downstream across domains	Unchecked submissions can shift checking work in the documented workflows; the cross-system examples do not establish a common net effect. The chapter's load-bearing account is science-scoped
Archive and models (Ch. 1)	Measured archive content, training-data use and recursive-training research were linked too closely	These facts do not establish that measured pages entered a particular corpus or caused present-day degradation
Marking duties (Ch. 4)	Text was described as the dominant medium of low-quality output, and jurisdictions were contrasted by whether they excluded it from marking duties	A comparison of specified actors, media, duties, exceptions and dates, without ranking jurisdictions
Peer review (Ch. 5)	A closed loop of AI-written papers, AI-written reviews and detector adjudication	At selected machine-learning venues, detector screening and estimated or documented AI-assisted reviewing coexist; the evidence does not establish an autonomous paper-review-decision cycle
Ad-fraud estimates (Ch. 5)	A multiplier was calculated across non-comparable vendor estimates	Differences in period, population, unit and extrapolation prevent treating these figures as a like-for-like loss range
Agent network (Ch. 2)	Registration counts were read as showing autonomous participation	Registrations do not establish activity or autonomy

TOPIC	WITHDRAWN CLAIM (NOT A FINDING)	WORDING IN FORCE
Other sectors (Ch. 5)	One uniform verification-cost mechanism operated across sectors and regions	The reviewed reports do not establish a common synthetic-content share, a uniform cost transfer or a net effect
UAE education (Ch. 6)	UAE measures were described as putting the tested intervention into practice, and the legal arrangements as exclusive of a labelling statute	Documented measures and stated objectives; implementation and outcomes are separate questions

7.3 What the evidence does not establish

- That web content measured as AI-associated is degrading current models.
- The complete chain from automated engagement, through learning, to the allocation of more fraudulent traffic.
- An autonomous paper-review-decision cycle at any venue.
- The total active population of the agent network in section 2.6, or the proportion operating autonomously. One study recorded 6,159 posting or commenting accounts within its specified early window^{37, 39}.
- A common synthetic-content share, or a uniform transfer of verification costs, across sectors.
- Several widely repeated figures that we could not trace to a qualifying source. They were excluded and are not repeated here.

7.4 What we searched and did not find

This review is not an exhaustive statement of AI-content obligations in the UAE, Saudi Arabia, Qatar, Bahrain or Oman. It records the instruments we identified and the limits of the source retrieval undertaken for this issue. In the materials we retrieved up to the close of the record, we did not locate a general duty covering all AI-generated output; specific media, election, advertising or sectoral rules may apply. For Chile, we did not verify a current obligation or measured phenomenon¹⁰². These are the results of our search, with its limits. They are not findings that no relevant measures exist.

7.5 What we located but did not clear

Official instruments from China, Japan, Brazil, Saudi Arabia, Egypt, Russia, Kazakhstan, Mexico, Argentina and Colombia, and a study from Morocco, were located. They were not cleared for detailed description, for reasons of retrieval, language or unresolved scope. Each appears in Figure 7 or in the references.

7.6 Sources: two different gaps

- **Ministry of Education 2026 classroom guide:** described through attributed reporting. The guide itself was not located.
- **Sharjah Private Education Authority:** the official announcement is verified. The underlying framework document was not retrieved.

7.7 Recorded discrepancies

- ICLR review and paper counts differ between press and primary sources²⁴.
- The title of the MoE guide varies in coverage¹⁰⁰.

- NeurIPS gives differing denominators (969 and 971)⁷. Separately, the mapping between its score-distribution results and its decision groups is not fully specified⁷.
- Different Moltbook counts come from different dates and populations^{37, 39}.
- French AI news-site counts use different frames²⁵.

Method

Sources. Every claim rests on a logged source, graded by type:

- original documents (laws, regulations, official announcements, peer-reviewed papers, court decisions);
- organisations' own reports and policies, always attributed in the sentence;
- claims reported by news outlets whose original we could not locate, attributed and marked as such;
- documents located but not verified in detail, shown only in tables with their status;
- claims we could not establish, recorded in Chapter 7.

Company figures are presented as what the company reports, not as independently validated measurements. Reports that rely on the same detection tool are not counted as independent confirmation of one another.

Verification and conflicts of interest. This issue was prepared with AI assistance. Because AI providers appear in it, material about each provider was checked by a model from a different provider. Every finding was tested to stand without the material about any single provider. No provider is presented as an example of good practice, and facts about individual providers appear only in Figure 6. An independent reviewer challenged each inference, and several were narrowed or withdrawn as a result (Chapter 7).

Pre-registration. Before verification began, we recorded in writing how each possible result would change the text, so that conclusions could not be fitted to findings afterwards. Interpretation Lock, SHA-256: f4915f763484ba8ddbe6ee2-fa95aec913992a772d8b79fb0114719855177a310. Amendment A1, SHA-256: 040ff0b0d-ca790d555702ca9043b291c410cff7d51c1a5d10ec8dbd08c006a22. The sealed documents are kept in the project archive; their fingerprints allow any later release to be checked against them.

Dates. The record closes on 18 October 2026, 23:59 (UTC+4). Later documents are not absorbed into this edition.

Figure data. The data behind each figure is published as CSV files with the online edition at noise.foresightflow.org.

Limitations.

- Most sources are in English. Official texts in other languages were described only where an English text or a verifiable passage was available.
- Arabic sources are paraphrased, not quoted.
- Estimates of AI-written content depend on detectors, which misclassify individual items.
- Chapter 6 reports announced education measures; it does not evaluate their implementation or outcomes.
- Legal and policy summaries are limited to the specified instruments and dates. They are not an exhaustive statement of applicable law, and not an assessment of anyone's compliance.

Corrections. info@foresightflow.org

References

Superscript numbers in the text link to these entries. Where an entry has no link, the source was identified but no stable public URL was recorded; organisations' own reports are attributed to them.

- 1 Pew Research Center (2026, 20 August). How much of the internet is written with AI? Data Labs, and accompanying methodology. <https://www.pewresearch.org/data-labs/2026/08/20/how-much-of-the-internet-is-written-with-ai/>
- 2 Turnitin. AI writing detection in the new, enhanced Similarity Report (guide). <https://guides.turnitin.com/hc/en-us/articles/22774058814093>
- 3 University of Waterloo (2025). Waterloo discontinuing the use of AI detection tool Turnitin.com: September 2025. <https://uwaterloo.ca/associate-vice-president-academic/news/waterloo-discontinuing-use-ai-detection-tool-turnitincom>
- 4 Curtin University (2025). Update on Turnitin AI detection tool (AI writing detection disabled from 1 January 2026). <https://www.curtin.edu.au/news/oasis-news/update-on-turnitin-ai-detection-tool/>
- 5 University of the Free State (2026, May). UFS shifts academic integrity approach in AI era (Turnitin AI detection discontinued from 1 July 2026). <https://www.ufs.ac.za/templates/news-archive/campus-news/2026/may/ufs-shifts-academic-integrity-approach-in-ai-era>
- 6 Matter of Newby v Adelphi University, 2026 NY Slip Op 26021 (Sup. Ct. Nassau County, 28 January 2026). https://www.nycourts.gov/reporter/3dseries/2026/2026_26021.htm
- 7 NeurIPS 2026 Program Chairs (2026, 2 June). AI-generated papers in the NeurIPS 2026 Position Paper Track. NeurIPS blog. <https://blog.neurips.cc/2026/06/02/ai-generated-papers-in-the-neurips-2026-position-paper-track/>
- 8 Liang, W., et al. (2024). Monitoring AI-modified content at scale: a case study on the impact of ChatGPT on AI conference peer reviews. Proceedings of ICML 2024, PMLR 235. <https://proceedings.mlr.press/v235/liang24b.html>
- 9 OpenAI (2026, 5 October). Our approach to EU text provenance rules. <https://openai.com/index/eu-text-provenance/>
- 10 Dathathri, S., et al. (2024). Scalable watermarking for identifying large language model outputs. Nature, 634, 818–823. <https://doi.org/10.1038/s41586-024-08025-4>; reference code: <https://github.com/google-deepmind/synthid-text>
- 11 arXiv (2026, 1 October). Updated rate limit policy. arXiv blog. <https://blog.arxiv.org/2026/10/01/updated-rate-limit-policy/>
- 12 Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: evidence from high school mathematics. PNAS, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- 13 Knowledge at Wharton. Without guardrails, generative AI can harm education. <https://knowledge.wharton.upenn.edu/article/without-guardrails-generative-ai-can-harm-education/>
- 14 Fan, Y., et al. (2025). Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>
- 15 Deezer (2026, 21 July). Newsroom release on fully AI-generated music uploads. <https://newsroom-deezer.com/2026/07/ai-music-exceeds-50-percent-daily-uploads-deezer/>
- 16 Deezer (2026, April). Newsroom release on AI-generated tracks as a share of new uploads. <https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>
- 17 Spotify (2025, 25 September). Spotify strengthens AI protections. Newsroom. <https://newsroom.spotify.com/2025-09-25/spotify-strengthens-ai-protections/>

- 18 Bandcamp (2026, 13 January). Keeping Bandcamp human. <https://blog.bandcamp.com/2026/01/13/keeping-bandcamp-human/>
- 19 Meta (2025, 25 September; 6 November). Introducing Vibes; Bringing Vibes to Europe. Meta Newsroom. <https://about.fb.com/news/2025/09/introducing-vibes-ai-videos/>
- 20 Bloomberg; TechCrunch; Axios (2025, 30 September). Reports on the release of the Sora social video app. <https://www.bloomberg.com/news/articles/2025-09-30/openai-releases-social-app-for-sharing-ai-videos-from-sora>
- 21 YouTube Help (2025, July). Channel monetisation policies: inauthentic content. <https://support.google.com/youtube/answer/1311392>
- 22 Kapwing (2025). AI slop report: the global rise of low-quality AI videos. <https://www.kapwing.com/blog/ai-slop-report-the-global-rise-of-low-quality-ai-videos/>
- 23 NewsGuard. AI Tracking Center: AI content farm news and information sites (update dated 23 June 2026). <https://www.newsguardtech.com/special-reports/ai-tracking-center/>
- 24 Pangram Labs (2025, 18 November). Pangram predicts 21% of ICLR reviews are AI-generated. <https://www.pangram.com/blog/pangram-predicts-21-of-iclr-reviews-are-ai-generated>
- 25 WAN-IFRA (2026, 10 July). He's found 15,000 fake AI news sites: here's what you need to know (interview with Jean-Marc Manach). <https://wan-ifra.org/2026/07/hes-found-15000-fake-ai-news-sites-heres-what-you-need-to-know/>
- 26 The Lancet (2026). Audit of fabricated references in published medical research, first seven weeks of 2026 (Columbia University). PII S0140-6736(26)00603-3. [https://doi.org/10.1016/S0140-6736\(26\)00603-3](https://doi.org/10.1016/S0140-6736(26)00603-3)
- 27 Baack, S. (2024, 6 February). Training data for the price of a sandwich: Common Crawl's impact on generative AI. Mozilla Foundation. <https://www.mozillafoundation.org/en/blog/Mozilla-Report-How-Common-Crawl-Data-Infrastructure-Shaped-the-Battle-Royale-over-Generative-AI/>
- 28 Common Crawl (2025). Submission to the UK's copyright and AI consultation. <https://commoncrawl.org/blog/submission-to-the-uks-copyright-and-ai-consultation>
- 29 Shumailov, I., et al. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- 30 Gerstgrasser, M., et al. (2024). Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. arXiv: 2404.01413. <https://arxiv.org/abs/2404.01413>
- 31 Hans, A., et al. (2024). Spotting LLMs with Binoculars: zero-shot detection of machine-generated text. Proceedings of ICML 2024, PMLR 235. <https://proceedings.mlr.press/v235/hans24a.html>
- 32 Graphite (2026). AI now writes as many online articles as humans do. <https://graphite.io/five-percent/research/ai-now-writes-as-many-online-articles-as-humans-do>
- 33 Originality.ai (2026, August). Dead internet tracker. <https://originality.ai/blog/dead-internet-tracker-august>
- 34 Emarat Al Youm (2026, 4 September). Report on the first week of the 2026–27 school year and AI use in UAE schools (Arabic). <https://www.emaratalyoum.com/local-section/education/2026-09-04-1.2079608>
- 35 Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7). arXiv: 2304.02819. <https://arxiv.org/abs/2304.02819>
- 36 Willison, S. (2026, 30 January). Moltbook. <https://simonwillison.net/2026/Jan/30/moltbook/>
- 37 Wiz Research (2026, 2 February). Exposed Moltbook database reveals millions of API keys. <https://www.wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys>
- 38 Reuters (2026, February). Interview with Ami Luttwak, Wiz, on the Moltbook exposure (as republished by Calcalist).
- 39 Holtz, D. (2026). The anatomy of the Moltbook social graph. arXiv:2602.10131. <https://arxiv.org/abs/2602.10131>

- 40 Claude Help Centre. How Claude marks AI-generated content (retrieved October 2026). <https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content>
- 41 OpenAI (2026, 19 May; update 31 July). Advancing content provenance. <https://openai.com/index/advancing-content-provenance/>
- 42 C2PA. Frequently asked questions. <https://c2pa.org/faqs/>
- 43 Saberi, M., et al. (2023). Robustness of AI-image detectors: fundamental limits and practical attacks. arXiv:2310.00076. <https://arxiv.org/abs/2310.00076>
- 44 CVPR 2026. Visual paraphrase attack on image watermarks (conference poster 39859). <https://cvpr.thecvf.com/virtual/2026/poster/39859>
- 45 Kazempourradi, S., Mojtahedi, R., & Mohseni, B. (2026). Watermarks and fingerprints as soft bindings for content provenance: an open-li-cence benchmark for images, audio and video. arXiv:2610.04151 (authors affiliated with Original Pictures Technologies, Inc.). <https://arxiv.org/abs/2610.04151>
- 46 House of Commons Library (2026, 20 January). AI content labelling (CBP-10467). <https://commonslibrary.parliament.uk/research-briefings/cbp-10467/>
- 47 Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. PNAS Nexus, 4(6), pgaf170. <https://doi.org/10.1093/pnasnexus/pgaf170>
- 48 National AI Centre, Department of Industry, Science and Resources (2025, November). Being clear about AI-generated content: a guide for business. <https://www.industry.gov.au/publications/being-clear-about-ai-generated-content>
- 49 Regulation (EU) 2024/1689 (Artificial Intelligence Act), Articles 50, 99 and 113, consolidated text as amended. AI Act Service Desk. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-50>
- 50 California AB 853 (2025), Chapter 674, Statutes of 2025. <https://legiscan.com/CA/text/AB853/2025>
- 51 European Commission (2026, 10 June). Commission publishes Code of Practice on marking and labelling of AI-generated content. <https://digital-strategy.ec.europa.eu/en/news/commission-publishes-code-practice-marking-and-labelling-ai-generated-content>
- 52 Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus). Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>
- 53 California SB 1000 (2026), Chapter 861, Statutes of 2026, approved 30 September 2026. <https://legiscan.com/CA/text/SB1000/2025>
- 54 California AB 2713 (2026). <https://legiscan.com/CA/text/AB2713/2025>
- 55 Ministry of Electronics and Information Technology, India (2026). Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, and FAQs on synthetically generated information. <https://www.meity.gov.in/>
- 56 Ministry of Science and ICT, Republic of Korea (2026). AI Basic Act commencement and guidance on marking of AI outputs. <https://www.msit.go.kr/eng/>
- 57 Official instruments located but not cleared for detailed description: Cyberspace Administration of China (cac.gov.cn); Diet of Japan (sangi-in.go.jp); Chamber of Deputies of Brazil (camara.leg.br); Superior Electoral Court of Brazil (tse.jus.br); SDAIA, Saudi Arabia (sdaia.gov.sa); ai.gov.eg (Egypt); publication.pravo.gov.ru (Russia); adilet.zan.kz (Kazakhstan); INE, Mexico (ine.mx); Public Prosecutor's Office, Argentina (fiscales.gob.ar); MinTIC, Colombia (mintic.gov.co); HACA, Morocco (haca.ma).
- 58 eSafety Commissioner (2024). Designated Internet Services Standard: registration fact sheet. <https://www.esafety.gov.au/sites/default/files/2024-06/Fact-sheet-registration-DIS-Standard.pdf>
- 59 WAM (2025, 25 September). UAE Media Council: using AI to depict national symbols without prior approval violates media content standards. <https://www.wam.ae/en/article/blvxe9y-uae-media-council-warns-against-misuse-harmful>
- 60 WAM (2026, 22 August). ICCMD unveils advanced digital monitoring system, launches International Charter for Media Integrity. <https://www.wam.ae/en/article/c1v24vf-iccmd-unveils-advanced-digital-monitoring-system>

- 61 Ministry of Communications and Information Technology, Qatar. Guidelines for secure adoption and usage of AI (ethical guidelines). <https://www.mcit.gov.qa/>
- 62 Information & eGovernment Authority, Bahrain (2025, July). General Policy for the Use of Artificial Intelligence. <https://www.iga.gov.bh/>
- 63 Ministry of Transport, Communications and Information Technology, Oman (2025, 1 April). General Policy for the Safe and Ethical Use of AI Systems. <https://www.mtcit.gov.om/>
- 64 Government of Viet Nam (2026). Law on Artificial Intelligence No. 134/2025/QH15 (English text). <https://en.baochinhphu.vn/viet-nams-law-on-artificial-intelligence-111260715113551787.htm>
- 65 Elections Department Singapore. Candidates' online election campaigning. <https://www.eld.gov.sg/candidate-parliamentary-campaign-online.html>
- 66 ANTARA (2025, 27 June). Indonesian govt, CSOs to develop tool to detect deepfakes. <https://en.antaranews.com/news/362473/indonesian-govt-csos-to-develop-tool-to-detect-deepfakes>
- 67 Ministry of Science, Technology and Innovation, Malaysia (2024). The National Guidelines on AI Governance and Ethics. <https://mas-tic.mosti.gov.my/storage/2024/09/THE-NATIONAL-GUIDELINES-ON-AI-GOVERNANCE-ETHICS.pdf>
- 68 Media Council of Kenya. Media guide on the use of artificial intelligence in Kenya. <https://mediacouncil.or.ke/>
- 69 National Council of Television and Radio Broadcasting of Ukraine. How to responsibly use artificial intelligence: guidelines developed for media. <https://webportal.nrada.gov.ua/en/how-to-responsibly-use-artificial-intelligence-guidelines-developed-for-media/>
- 70 NCC-CSIRT, Nigerian Communications Commission (2026). Advisory 2026-026: AI-driven social engineering and deepfake threats. <https://csirt.ncc.gov.ng/>
- 71 arXiv (2025, 31 October; 2026, 21 January). Policy updates on review and position papers in computer science, and on endorsement. arXiv blog. <https://blog.arxiv.org/2026/01/21/attention-authors-updated-endorsement-policy/>
- 72 Research Information (2026). arXiv imposes one-year ban for unchecked AI-generated content. <https://www.researchinformation.info/news/arxiv-imposes-one-year-ban-for-unchecked-ai-generated-content/>
- 73 CVPR, ICCV, AAAI, WSDM and KDD (2025–2026). Author guidelines and submission policies for the respective conference editions and tracks.
- 74 ICLR (2025, 15 April). Leveraging LLM feedback to enhance review quality. ICLR blog; accompanying paper arXiv:2504.09737. <https://blog.iclr.cc/2025/04/15/leveraging-llm-feedback-to-enhance-review-quality/>
- 75 Robert Half (2026, 14 September). Robert Half research reveals AI-generated applications are slowing hiring (UK). <https://press-releases.responsesource.com/news/107598/>
- 76 Robert Half (2026, 10 March). Robert Half survey: 67% of HR leaders report AI-generated applications are slowing hiring (US). <https://press.roberthalf.com/2026-03-10-Robert-Half-survey-67-of-HR-leaders-report-AI-generated-applications-are-slowing-hiring>
- 77 Robert Walters (2026). Middle East employer survey on AI-assisted job applications (company post on LinkedIn; link is to coverage by MIT Sloan Management Review Middle East). <https://www.mitsloanme.com/article/42-of-middle-east-employers-are-drowning-in-ai-generated-job-applications-survey-reveals/>
- 78 Amazon Kindle Direct Publishing. Content guidelines; title creation limits (help pages retrieved October 2026). https://kdp.amazon.com/en_US/help/topic/6200672390
- 79 Federal Trade Commission (2024, August). Final rule banning fake reviews and testimonials (16 CFR Part 465); Consumer Reviews and Testimonials Rule: questions and answers; cases and proceedings (2026). <https://www.ftc.gov/business-guidance/resources/consumer-reviews-testimonials-rule-questions-answers>
- 80 Federal Trade Commission (2025, 22 December). FTC reopens and sets aside Rytr final order. <https://www.ftc.gov/news-events/news/press-releases/2025/12/ftc-reopens-sets-aside-rytr-final-order-response-trump-administrations-ai-action-plan>
- 81 Pangram Labs (2026, May). AI-generated Amazon reviews. <https://www.pangram.com/blog/ai-amazon-reviews>

- 82 Thales / Imperva (2026). Bad Bot Report 2026. <https://www.imperva.com/blog/bad-bot-report-2026-bots-agentic-age/>
- 83 HUMAN Security (2026). 2026 State of AI Traffic and Cyberthreat Benchmarks. <https://www.humansecurity.com/learn/resources/2026-state-of-ai-traffic-cyberthreat-benchmarks/>
- 84 Spider Labs (2026, 2 September). Spider AF report reveals \$25.3B global first-half ad fraud losses (press release). <https://www.news-wire.com/news/spider-af-report-reveals-25-3b-global-first-half-ad-fraud-losses-22855835>
- 85 Fraudlogix (2026). Ad fraud statistics 2026. <https://www.fraudlogix.com/stats/ad-fraud-statistics-2026/>
- 86 DoubleVerify (2026, 7 May). Global study: fuelled by AI, CTV fraud schemes surge 140% globally. <https://doubleverify.com/company/news-room/global-study-ctv-fraud-schemes-surge-140-globally>
- 87 Immorlica, N., Jain, K., Mahdian, M., & Talwar, K. (2005). Click fraud resistant methods for learning click-through rates. WINE 2005, LNCS 3828, 34–45. https://doi.org/10.1007/11600930_5
- 88 SIGIR 2026. Paper on post-click conversion-rate prediction with noisy conversion labels. <https://doi.org/10.1145/3805712.3809570>
- 89 OECD (2026). OECD Digital Education Outlook 2026. OECD Publishing.
- 90 Pew Research Center (2026, 24 February). How teens use and view AI. <https://www.pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai/>
- 91 Higher Education Policy Institute (2026). Student Generative AI Survey 2026 (HEPI Report 199). <https://www.hepi.ac.uk/wp-content/uploads/2026/03/HEPI-Report-199-Gen-AI-Survey-2026.pdf>
- 92 Lee, H.-P., et al. (2025). The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. Proceedings of CHI 2025. <https://doi.org/10.1145/3706598.3713778>
- 93 Kosmyrna, N., et al. (2025). Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv: 2506.08872 (preprint). <https://arxiv.org/abs/2506.08872>
- 94 OECD (2024, December). Do adults have the skills they need to thrive in a changing world? Survey of Adult Skills 2023. https://www.oecd.org/en/publications/do-adults-have-the-skills-they-need-to-thrive-in-a-changing-world_b263dc5d-en.html
- 95 Ministry of Education, UAE (2025). Ministry of Education introduces AI curriculum in public schools starting from the 2025–2026 academic year (press release; WAM). <https://www.wam.ae/en/article/bji6z2n-ministry-education-introduces-curriculum-public>
- 96 Abu Dhabi Department of Education and Knowledge (2026, 17–18 September). ADEK advances AI Literacy Initiative across 170 private schools. <https://www.adek.gov.ae/Media-Centre/News/ADEK-Advances-AI-Literacy-Initiative-across-170-Private-Schools>
- 97 WAM (2026, 14 October). Sharjah Private Education Authority launches ethical framework for AI use in schools. <https://www.wam.ae/en/article/c2qm385-sharjah-private-education-authority-launches>
- 98 Khaleej Times (2025, August). AI in RAK schools: teachers, curriculum (statements by the Ras Al Khaimah Department of Knowledge). <https://www.khaleejtimes.com/uae/ai-in-rak-schools-teachers-curriculum>
- 99 Khaleej Times; The National (2026, 1 October). Statements by KHDA on a framework for AI and technology use in Dubai schools. <https://www.khaleejtimes.com/uae/education/khda-set-clear-direction-ai-technology-use-dubai-schools>
- 100 Gulf News (2026, February). UAE sets strict new rules on generative AI in schools, banning use for under-13s; Education Middle East (2026). UAE issues new AI guidelines for schools. <https://gulfnnews.com/uae/education/uae-sets-strict-new-rules-on-generative-ai-in-schools-banning-use-for-under-13s-1.500443163>
- 101 Rest of World (2026, 13 October). Report on AI education in UAE schools. <https://restofworld.org/2026/uae-mandates-k12-ai-education-schools/>
- 102 Ministerio de Ciencia, Tecnología, Conocimiento e Innovación, Chile. Account of the AI bill's legislative progress. <https://www.mincien-cia.gob.cl/>

- 103 WAM (2025, 18 December). UAE Government issues Federal Decree-Law establishing, organising National Media Authority; National Media Authority, FAQ (Federal Decree-Law No. 11 of 2025). <https://www.wam.ae/en/article/bn9xsue-uae-government-issues-federal-decree-law>
-
- 104 Khaleej Times (2026, 30 September). UAE National Media Authority warns against AI images of national symbols; Gulf News (2026, 29 September). UAE warns against AI-generated images of national symbols, public figures. The Authority's post itself was not retrieved. <https://www.khaleejtimes.com/uae/uae-national-media-authority-warns-against-ai-images-of-national-symbols-rules>
-
- 105 European Commission (2026, 20 July). Guidelines on transparency obligations for providers and deployers of AI systems. <https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems>
-

Not an allegation against any named person or body. Nothing here constitutes legal advice. Statements of fact are limited to matters recorded in the public sources cited and are reports of what those sources state.