

CASOVER

AI and authors

Research and evidence
Craft, factual reliability and publishing decisions

Commissioned by Steve Castledine, Draft2Book LLP

Edited research companion | 6 October 2026

Research edition 0.2 | Selected-source review

Contents

About this companion	3
1. AI and authors: the research narrative	4
2. Citation accuracy, retrieval and attribution	9
3. Human trust, checking and error detection	18
4. Craft, voice, cultural expression and skills	23
5. Detectors and publishing policies	27
6. Rights, confidentiality and legal-source boundaries	29
7. Proposition map and boundaries	31
8. Unresolved evidence and future research	34
9. References and original-source links	36

Use the PDF bookmarks or linked source labels to navigate.

About this companion

Commissioned by Steve Castledine, Draft2Book LLP. Prepared for Casover, for fiction and non-fiction authors, editors and publishers.

Research discovery used Perplexity Deep Research. Selected original-source checks, synthesis and preparation used OpenAI Codex under Steve Castledine's direction. This PDF is an edited research companion to the Casover Insight, not a new experiment or a systematic review of all literature.

Research cut-off: 6 October 2026. Research edition 0.2. The original Notion research and generated reports have been preserved. This edition removes private planning, navigation, prompts, progress comments and instructions, and presents the substantive narrative, all 23 evidence-register records, craft source reviews, detector and policy checks, legal-source notes and claim boundaries.

The six generated research reports were discovery material. Their complete interface source panels were not captured, and their verification labels are not independent approval. This companion uses the separate evidence records and selected original-source checks; it does not republish unchecked report claims as established findings. Additional leads are listed separately.

How to read it

Read the narrative first for the argument and practical implications. Use the detailed evidence sections to inspect the task, model, finding and boundary behind a claim. The proposition map provides a compact route between the argument and the evidence. The final references link to original publications, official pages or clearly identified author manuscripts.

What a recorded source check means

A recorded check means the specified source passage was inspected with AI assistance in the research workflow on 6 October 2026. It does not mean that all results were reproduced, supplements reviewed, publication versions reconciled or correction notices exhaustively searched. The research record retains these limits. PDF preparation has not silently upgraded any source's status.

Selected-text support means the recorded passage supports the narrowly stated finding. **Partial support** means an important part of the checking remains incomplete. **Discovery lead** means the evidence is not adopted as a substantive finding. These describe inspection coverage, not certainty or a universal quality score.

Independent human reading of priority originals remains outstanding. Legal materials are presented as source notes and questions, with their status identified. This companion makes no definitive current-law determination and has not been endorsed by any cited researcher, organisation or law firm.

1. AI and authors: the research narrative

Executive summary

Authors need to assess what AI actually does in their work. Suggesting questions, locating possible sources, offering a critique, rewriting a paragraph and producing a complete draft involve different creative choices and checks. Describing all of these simply as “AI-written” obscures the decisions that authors and publishers need to make.

The selected research supports a conditional position. AI assistance can improve particular writing outcomes, yet benefits can coexist with reduced diversity, weak understanding, factual errors or loss of control. Short-task results do not settle what happens across a complete novel, a substantial non-fiction book or years of professional practice. An AI detector measures patterns under particular conditions; its score does not establish the whole history of a manuscript.

Our proposed approach is human-led and accountable: choose a defined purpose, protect the material entrusted to the author, examine suggestions against creative intent, verify factual claims in original sources, retain a proportionate record of significant decisions and check the applicable publishing terms. These are editorial recommendations. Their complete effectiveness has not been experimentally established.

This paper is a working reference for fiction and non-fiction authors, editors and publishers, with UK practice as its context. It separates empirical findings, editorial interpretation, stated organisational policies and legal questions. It does not yet establish current-law conclusions, industry-wide detector use or a general improvement rate since 2023.

1. Describe the activity before judging the result

The first question is practical: what assistance was requested, what material was supplied and what was accepted? An author might ask for questions about a character's motivation, alternative chapter sequences, possible background sources, grammatical suggestions or replacement prose. Each use needs a description accurate enough for an editor, reader or contracting party to understand.

For this paper, “AI-assisted” describes a process in which a person uses an AI tool while making substantive decisions about the work. “AI-generated” describes material produced by the tool. A manuscript can contain both. These working descriptions do not override legal tests or the definitions in a particular contract, submission policy or platform rule.

Evaluation should consider the purpose, the nature of generated expression, the author's scrutiny, factual support, input sensitivity and applicable terms. A polished paragraph might be factually wrong; an original human passage might breach a submission rule if it has been edited in a prohibited way. Writing quality, copyright, infringement, disclosure and provenance require their own enquiries.

2. Craft, voice and creative judgement

Doshi and Hauser's 2024 experiment found that access to GPT-4 story ideas improved ratings of brief stories, especially for participants with lower baseline divergent-association scores. Stories within the AI-assisted conditions also became more similar. The trial involved 293 UK participants writing eight-sentence stories and 600 evaluators, rather than professional novelists completing books. It shows that individual ratings and collective diversity can move in different directions. [Original author manuscript](#), main results and methods; C01.

Padmakumar and He's ICLR 2024 study adds an important qualification: its diversity results differed between GPT-3 and instruction-tuned InstructGPT. Thirty-eight writers produced 300 short essays across conditions. The measured similarity effect was not the same for both tools. We should therefore avoid treating all forms of assistance, or all models, as interchangeable. [Original conference paper](#), sections 2-6; C02.

Professional accounts help explain how these tensions arise in practice. Ippolito and colleagues commissioned 13 published writers to use LaMDA-based Wordcraft for eight weeks on short stories. Participants explored useful workflows while encountering problems with voice, repetition and understanding. This was a small developer-led qualitative study, without a randomised control group. [Author manuscript](#), sections 3-4, 6-7 and 9; A006-S01.

Guo and colleagues observed 18 existing AI-using creative writers. Their accounts show deliberate limits on assistance to protect voice, enjoyment and a feeling of ownership; they do not demonstrate causal preservation of craft. [Author manuscript v4](#), sections 3, 4.1 and 6; A006-S04.

Agarwal and colleagues' controlled study found cultural convergence in brief English tasks with GPT-4o autocomplete. It raises a specific concern about cultural expression, while country proxies, selected participants and text-comparison methods limit generalisation. [Author manuscript v3](#), sections 3, 4.3 and 5.4; A006-S06.

The editorial implication is to judge suggestions against the author's intended effect. Does a revision improve a character's motivation, preserve a deliberate ambiguity or clarify a transition? Or does it smooth away the diction, cultural detail or uncertainty that gives the work its character? Asking for reasons and alternatives may help an author examine those choices, but this is proposed practice rather than proof that a particular prompting method preserves voice.

Learning needs its own assessment. Better text produced with assistance does not establish better independent writing afterwards. The evidence checked for this draft does not settle long-term skill retention, modern book-length editing or the full range of accessibility needs. Authors should have room to choose assistance or decline it. Research into queer and disabled fiction writers records serious concerns about precarity; it does not demonstrate universal assistive benefit or harm. [Lamb and colleagues](#), methods and limitations; A006-S03.

One experiment found better immediate unaided product-design performance after guided AI use than after unrestricted use. It involved Singapore students, not authors; no equally guided human-only comparison or delayed follow-up isolated lasting benefits. [Wong and Qiu](#), Method, Results and Limitations; A006-S05.

3. Factual research and citations

Fiction can invent its events while still depending on real details about places, history, occupations or medicine. Non-fiction may build its argument directly on cited evidence. In both cases, an author needs to distinguish a plausible suggestion from a checked fact.

Four questions should be kept separate: does the source exist; are its bibliographic details right; is it relevant; and does the original support the precise claim? Walters and Wilder's 2023 comparison documented fabricated references and errors in real references. Its model comparison is historical, not a present-day accuracy estimate. [Original study](#), Methods and Tables 3-4; E001.

A more recent study by Linardon and colleagues examined 333 citation-claims in twelve mental-health literature reviews generated in December 2025 using the authors' reported ChatGPT-5 thinking model. All citations were traceable to identifiable publications, yet claims could misrepresent their content. The scoring conservatively retained the worst claim within a cited passage. This domain-specific test does not provide a general error rate for authors' work. [Original study](#), Methods and Results; E007.

The practical recommendation is to open the original, read the supporting passage and its qualifications, and retain the location behind any important assertion. A DOI check establishes identity, not support. Search access can supply evidence but does not by itself establish accurate interpretation. Repeating a claim through several AI answers also does not establish independent corroboration.

Source checking for this project identified generated links that pointed to different papers from the studies being discussed. Corrected originals are used here. The error illustrates why a working link and a plausible title require checking against the precise claim.

4. Human checking and reliance

AI explanations can help people, and can also be persuasive when wrong. Si and colleagues found an overall benefit in a selected fact-checking task alongside worse decisions for a subset with incorrect explanations. Both findings belong in the interpretation. This was a historical model and a difficult claim-selection design; a sample-arithmetic inconsistency still requires reconciliation. [Original publication](#), sections 4-5; E018.

For authors, checking should therefore involve evidence beyond the answer's fluency or confidence. Ask what would disprove the assertion, inspect contrary evidence and distinguish opening a source from reading and understanding it. Where the consequences of an error are substantial, obtain appropriate subject expertise.

This is a proportionate editorial approach. It does not imply that every sentence needs the same investigation or that every person invariably over-trusts AI. Record what was actually checked and unresolved, rather than using "verified" as a general badge for a report.

5. Rights, contracts and confidentiality

The UK's historical official explanation distinguishes human work made with AI assistance from computer-generated work without a human author. The March 2026 government report also discusses protection for human creative expression and proposals concerning computer-generated works. These official documents explain policy and legal issues; they are not themselves amendments to copyright law. [IPO consultation document](#), Existing copyright protection; [2026 policy report](#), Section I.

Lewis Silkin's briefing, displayed as published on 1 October 2026, recommends meaningful human authorship, documentation, output clearance and attention to indemnity and insurance. It supports an accountable creative process, without guaranteeing copyright or freedom from infringement. Its named authors are Oliver Fairhurst and Adrian Aronsson-Storrier. [Briefing](#), final practical-advice section.

The proposed author record should explain significant creative choices, the assistance used and subsequent revisions. It can help a reviewer understand the process, but does not automatically prove ownership, policy compliance or absence of copied expression. Before supplying unpublished third-party material or personal information to a service, establish authority and examine relevant confidentiality, data-use and contractual terms.

Current amended legislation, later developments and application to particular manuscripts remain for qualified legal review. Questions about training material, output infringement, an author's own copyright and contractual disclosure should be assessed separately.

6. Publishing policies and detector decisions

Publishing organisations have different stated rules. Greene & Heaton's live submission page rejects AI-originated, written or edited submissions, including supporting submission material. The Book Dragon Publishing Press distinguishes assistance from generation and requires disclosure. These two pages show contrasting approaches; they do not establish the practice of the whole industry. [Agency policy](#); [publisher policy](#); A005-S04-S05.

A detector result should be interpreted using its build, threshold, test population, passage length, language, genre and treatment of mixed writing. Jaeho Lee's July 2026 independent evaluation found strong performance under some conditions and more missed generations after style-imitation prompting. It tested short passages, not manuscript histories or complete novels. Zero observed false positives in a sample does not establish zero risk in other writing. [Original evaluation](#), data, results and limitations; A005-S01.

The evidence checked here supports neither blanket dismissal of detectors nor treating their scores as conclusive proof of who wrote a book. We have not established how often UK trade publishers rely on detectors, or a representative set of sanctions against authors.

Our recommendation is that an adverse assessment should identify the applicable rule, explain the evidence and its limits, examine relevant drafts and allow a meaningful author response. That is a proposed fair process, not a claim that every author has an enforceable right to publication or appeal. Authors should describe use honestly and address the actual policy rather than attempt to manipulate a detector.

7. A practical author workflow

The following is proposed editorial practice, informed by the selected evidence:

1. **Set the purpose.** Define the task and what decisions remain with the author. Save the starting draft.
2. **Check permission and terms.** Establish whether the material can be supplied and whether the intended use is permitted.
3. **Keep creative intent explicit.** Ask for options or critique where useful. Compare suggestions with the intended voice, meaning and reader experience.
4. **Make the decision.** Accept, adapt or reject suggestions consciously. Retain a short explanation for significant changes.
5. **Check factual support.** Open important originals, record exact locations and qualify unsupported assertions.
6. **Review the whole work.** Check continuity, consistency and consequences across chapters. Seek human editorial or specialist input where appropriate.
7. **Prepare the submission record.** Recheck the actual terms, disclose as required and keep proportionate drafts, revisions and source notes.

For example, a historical novelist could ask for questions about a scene, then research the equipment and chronology independently. A non-fiction author could use a tool to identify possible objections, then check the original studies before incorporating them. These are illustrative applications, not measured outcomes.

Unwise use includes publishing unexamined factual output, citing an unread source as authority, accepting substantial changes without understanding them, or concealing assistance contrary to the applicable terms. The concern is the decision and its consequences, rather than a tool's name alone.

8. Research limitations and open questions

This companion is a selective synthesis, not a systematic review or a legally approved reference. Priority work includes modern professional book studies, learning after AI removal, multilingual and accessibility evidence, inaccessible methods, correction and publication-version checks, primary records of disputed decisions and current-law review.

The reference should remain dated and correctable. Changes to a model, publisher policy or legal position should trigger a relevant update, with the previous wording preserved. Priority originals still require independent human inspection; legal conclusions require qualified review.

The central proposition is straightforward: assess authors' AI use through the actual creative process, the reliability of the resulting material and the applicable obligations. That approach accommodates useful assistance while providing a clear basis for identifying misuse and assessing disputed publishing decisions.

2. Citation accuracy, retrieval and attribution

The first fifteen register records examine reference existence, bibliographic accuracy, claim support and retrieval. Their tasks and scoring rules differ. Percentages across them should not be combined into an overall AI accuracy rate. Records with unresolved table or summary discrepancies retain those qualifications.

E001 | Fabrication and errors in the bibliographic citations generated by ChatGPT

Authors: William H Walters; Esther Isabelle Wilder

Model and test date: GPT-3.5 and GPT-4; first week April 2023.

Design and sample: 84 papers on 42 topics; 636 references.

What was measured: Existence through title/author matching; substantive bibliographic errors assessed separately.

Recorded finding: Fabrication: 55% GPT-3.5, 18% GPT-4; substantive errors among real citations: 43% and 24%.

Inspected location: Methods; Tables 3 and 4.

Limits: Historical prompts and models; cannot estimate October 2026 performance.

Check coverage: Selected-text support. No notice identified in inspected repository article; dedicated publisher/Crossmark and retraction checks outstanding.

Provisional research rating: Moderate. Direct comparative empirical evidence within the tested conditions.

[Original source](#)

Identifier: 10.1038/s41598-023-41032-5.

E002 | Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis

Authors: Mikaël Chelli and colleagues

Model and test date: 2023 versions of GPT-3.5, GPT-4 and Bard; PubMed baseline July 2023.

Design and sample: 11 rotator-cuff systematic reviews; 33 prompts; 362 generated references plus 109 baseline references.

What was measured: Hallucination if at least two of title, first author and publication year incorrect.

Recorded finding: Hallucination: GPT-3.5 39.6%, GPT-4 28.6%, Bard 91.4%.

Inspected location: Methods; Results; Table 3; Figure 3.

Limits: Specialist domain, selected prompt variants, historical models. Eligibility is not sentence-level claim entailment.

Check coverage: Selected-text support. No notice identified in inspected repository article; dedicated publisher/Crossmark and retraction checks outstanding.

Provisional research rating: Moderate. Direct comparison with human systematic reviews; narrow domain.

[Original source](#)

Identifier: 10.2196/53164.

E003 | Accuracy of Chatbots in Citing Journal Articles

Authors: Anjun Chen; Drake O Chen

Model and test date: GPT-4 and default GPT-3.5; 20 April to 6 May 2023.

Design and sample: Cross-sectional comparison; 419 references on learning-health-system topics.

What was measured: Checking cited journal and Google Scholar; comparison of bibliographic fields; failed verification classified fake.

Recorded finding: 159/162 GPT-3.5 references classified fake (98.1%); GPT-4 53/257 (20.6%).

Inspected location: Author header; Methods; Results; Table 2.

Limits: Narrow healthcare context and historical versions. Broad failure definition; not directly comparable with other studies.

Qualifications: The correct co-author name is Drake O Chen.

Check coverage: Selected-text support. No notice identified in inspected repository article; dedicated publisher/Crossmark and retraction checks outstanding.

Provisional research rating: Moderate. Original comparative evidence; scope and error definition limit generalisation.

[Original source](#)

Identifier: 10.1001/jamanetworkopen.2023.27647.

E004 | Hallucination Rate of Peer-Reviewed Citations Generated by Large Language Models in Neurocritical Care

Authors: Ali Seifi; Ali Seyfi

Model and test date: 10 March 2026: GPT-5.3, DeepSeek-V3, Grok-4; paid web interfaces; retrieval disabled; one response per topic.

Design and sample: 300 references; ten neurocritical-care topics; two blinded expert raters.

What was measured: Any bibliographic inaccuracy differs from complete fabrication.

Recorded finding: Any inaccuracy: 165/300. Fabrication: 85/300. Per-model fabrication: GPT 27/100; DeepSeek 8/100; Grok 50/100.

Inspected location: Methods; Results; Table 1; Limitations.

Limits: Specialist domain; single run; uncontrolled backend routing. No 2023 comparator or retrieval-enabled arm.

Qualifications: Published 25 August 2026. Model names are authors' reported labels.

Check coverage: Selected-text support. Repository inspected; no notice identified. Publisher/Crossmark check pending.

Provisional research rating: Moderate. Blinded comparative test; limited sampling.

[Original source](#)

Identifier: 10.1097/CCE.0000000000001474.

E005 | Synthesizing scientific literature with retrieval-augmented language models

Authors: Akari Asai and colleagues

Model and test date: Includes gpt-4o-2024-05-13, Llama 3.1 and OpenScholar; exact execution dates pending.

Design and sample: ScholarQABench; automatic scoring and expert evaluation.

What was measured: Title existence and claim support assessed separately.

Recorded finding: Table 2 reports GPT-4o title fabrication 78.7% (computer science), 94.8% (biomedicine); OpenScholar-8B zero in that analysis.

Inspected location: Baseline models; Table 2; Methods; Limitations; Change history.

Limits: Developer-led specialist benchmark; no general 2023-2026 trend estimate.

Qualifications: Published 4 February 2026. Abstract's 78-90% differs from Table 2's biomedicine rate; reconcile before quotation. Dataset totals differ between sections.

Check coverage: Partial support. Publisher change history: 11 March 2026 acknowledgement addition for DARPA funding.

Provisional research rating: Moderate. Direct comparisons; supplementary details still require review.

[Original source](#)

Identifier: 10.1038/s41586-025-10072-4.

E006 | Citation Constraints and Reference Hallucinations in Large Language Models

Authors: Kimberly Davis; Qusay Mahmoud

Model and test date: Gemini 3; ChatGPT 5.1; ChatGPT 4o; Microsoft 365 Copilot. Test dates/settings not yet inspected.

Design and sample: Abstract reports 36 academic-style documents across four systems.

Recorded finding: Abstract reports more invalid or inconsistent references with stricter requirements.

Qualifications: PDF tool download timed out. Treat as a priority lead, not an audited empirical conclusion.

Check coverage: Discovery lead; not adopted. Not completed.

Provisional research rating: Unverified. Publisher metadata and abstract confirmed; full methods unavailable during this pass.

[Original source](#)

Identifier: PMLR 318:464-475; davis26a.

E007 | Fluency Without Fidelity: Errors in Citation-Attributed Claims in Large Language Model-Generated Literature Reviews in Mental Health

Authors: Jake Linardon; Mariel Messer; Cleo Anderson; Olivia Marie Soliman; Claudia Liu; Joseph Firth; John Torous

Inspected location: Methods: Accuracy Scoring; Results

Limits: One mental-health test; supplements and browsing configuration unchecked.

Design, model and date: Twelve mental-health literature reviews generated in December 2025, using the authors' reported ChatGPT-5 thinking model; 333 citation-claims linked to 209 identifiable publications. Published April 2026.

Scoring qualification: All citations were identifiable; 145/333 citation-claims were judged completely accurate. A passage with multiple claims retained its lowest claim score. This is a domain-specific test, not a general author error rate.

Qualifications: The generated discovery link was corrected. The corresponding original was checked separately.

Check coverage: Partial support. Dedicated correction and retraction check pending

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.1007/s41347-026-00630-6.

E008 | Large Language Models Require Curated Context for Reliable Political Fact-Checking; Even with Reasoning and Web Search

Authors: Matthew R. DeVerna; Kai-Cheng Yang; Harry Yaojun Yan; Filippo Menczer

Design and sample: 6,137 political claims; macro F1 comparison

Inspected location: Author manuscript, Table A2, page 16

Limits: Curated context contains summaries of existing fact-checks. Not a universal citation-accuracy test.

Recorded finding: Table A2 reports GPT-4o macro F1 values on 6,137 political claims: 0.26 without retrieval, 0.76 with web search, 0.90 with curated retrieval and 0.92 with both. Macro F1 summarises classification performance; it is not a citation-error percentage. Curated context includes summaries of existing fact-checks.

Qualifications: Exact test dates, publication version, identifier and correction status remain incomplete. The generated discovery link was corrected to the author manuscript.

Check coverage: Partial support. Dedicated correction and retraction check pending

Provisional research rating: Moderate.

[Original source](#)

E009 | This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance

Authors: Courtni Byun; Piper Vasicek; Kevin Seppi

Model and test date: text-davinci-003, GPT-3.5-turbo, GPT-4; exact snapshots/run dates unchecked. Temperature 0.

Design and sample: 20 CHI/EMNLP papers; 1,616 annotated citations after excluding missing titles.

What was measured: Exact-title search on Google Scholar's first page; relevance uses target-bibliography overlap.

Recorded finding: Table 1 supports GPT-4 title accuracy 66.00%, 72.22%, 37.50% versus GPT-3.5 42.00%, 30.39%, 22.38%.

Inspected location: Sections 3-4; Tables 1-2; Appendix A

Limits: Restricted historical tasks; absence from one search page is not conclusive nonexistence. No general 2026 inference.

Check coverage: Partial support. Anthology record inspected; formal correction/retraction search pending.

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.18653/v1/2024.hcinlp-1.3.

E010 | Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties

Authors: Yasin Celal Güneş; Turay Cesur; Eren Çamur

Model and test date: Six reported product variants; December 2024; default settings.

Design and sample: 120 questions; 480 references per model, supported by Table 2.

What was measured: Composite bibliographic/content score, not separate atomic claim entailment.

Recorded finding: Claude 3.5 Sonnet: fabricated 15/480; fully accurate 388/480. Gemini 1.5 Pro: 291/480 and 13/480.

Inspected location: Methods; Table 2

Limits: Single outputs; one blinded scorer. RAS wording/count summaries need reconciliation.

Qualifications: Epub 12 May 2025; issue March 2026. The test date is distinct from publication and issue dates.

Check coverage: Partial support. Published methodological letter and author reply inspected; formal notices check pending.

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.4274/dir.2025.253101.

E011 | Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation

Authors: Sirui Xia; Xintao Wang; Jiaqing Liang; Yifei Zhang; Weikang Zhou; Jiaji Deng; Fei Yu; Yanghua Xiao

Model and test date: Llama3-8B main comparison; exact run dates/snapshot reconciliation pending.

Design and sample: ASQA 948, ELI5 1,000, EXPERTQA 1,000; fine-tuned interleaved generation.

What was measured: Automated NLI entailment; answer correctness uses recall metrics.

Recorded finding: Table 2 CAS: 92.1, 89.9, 90.1; mean 90.7. Authors report 31.3% relative gain and 6.0% lower answer accuracy versus ALCE/ChatGPT.

Inspected location: Sections 4.1-4.2, Table 2; Appendix C.3

Limits: Oracle passages for ASQA/ELI5; differing models. Not a human-verified error rate or universal retrieval effect.

Check coverage: Partial support. Anthology record inspected; dedicated notices check pending.

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.18653/v1/2025.findings-naacl.55.

E012 | Influence of Topic Familiarity and Prompt Specificity on Citation Fabrication in Mental Health Research Using Large Language Models: Experimental Study

Authors: Jake Linardon; Hannah K Jarman; Zoe McClure; Cleo Anderson; Claudia Liu; Mariel Messer

Model and test date: GPT-4o; June 2025

Design and sample: Six generated reviews; 176 references. Database and publisher searches. No human participants.

What was measured: Reference existence and bibliographic correctness; no semantic claim-support score.

Recorded finding: Table 1: 35/176 fabricated; 64/141 genuine references contained errors. Overall general/specialised fabrication difference was not significant ($P=.21$).

Inspected location: Methods: Data Collection and Analysis; Table 1; Fabrication Rates; Accuracy Rates.

Limits: Six outputs; topic characteristics confounded; exact model snapshot and browsing settings unclear.

Qualifications: Abstract BDD accuracy differs from Table 1 (20/34 versus 10/34). Abstract 'nearly two-thirds' conflicts with 99/176=56.25%.

Check coverage: Partial support. Original inspected; dedicated notices and identifier resolution remain unchecked.

Provisional research rating: Limited.

[Original source](#)

Identifier: 10.2196/80371.

E013 | CiteFix: Enhancing RAG Accuracy Through Post-Processing Citation Correction

Authors: Harsh Maheshwari; Srikanth Tenneti; Alwarappan Nakkiran

Model and test date: Anonymised proprietary Models A/B/C; Qwen variants; exact test dates pending.

Design and sample: Table 2: Model C correction comparison, 50 representative questions; two expert reviewers.

What was measured: MQLA: question passes multiple quality thresholds; combined metric, not citation accuracy alone.

Recorded finding: Citation post-processing improved the reported combined quality metric under tested conditions.

Inspected location: Sections 3-4; Table 2.

Limits: Small proprietary evaluation; models anonymised. Do not interpret relative MQLA improvement as percentage-point citation-error reduction.

Qualifications: Abstract 15.46% relative improvement requires reconciliation with rounded Table 2 entries, including 15.8%. Dedicated notices/data checks pending.

Check coverage: Partial support. Original inspected; dedicated notices and identifier resolution remain unchecked.

Provisional research rating: Limited.

[Original source](#)

Identifier: 10.18653/v1/2025.acl-industry.23.

E014 | Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents

Authors: Hailey Onweller; Elias Lumer; Austin Huber; Pia Ramchandani; Vamse Kumar Subbiah; Corey Feld

Model and test date: Named frontier/search systems; exact snapshots and test dates pending.

Design and sample: Preprint v1; 14 systems, 130 research queries; automated relevance/fact-check judges, manual calibration.

What was measured: Link validity, relevance and citation support are separate measures.

Recorded finding: Table 1 illustrates that valid links can coexist with substantially lower automated support scores.

Inspected location: HTML v1 Methods and Table 1.

Limits: Preprint; automated judging, truncation to 5,000 characters, denominator and calibration detail need checks. Not a human-trust experiment.

Qualifications: Submitted 7 May 2026. Peer review not established; current-version and notices checks pending.

Check coverage: Partial support. Original inspected; dedicated notices and identifier resolution remain unchecked.

Provisional research rating: Limited.

[Original source](#)

Identifier: 10.48550/arXiv.2605.06635.

E015 | The Attribution Crisis in LLM Search Results

Authors: Ilan Strauss; Jangho Yang; Tim O'Reilly; Sruly Rosenblat; Isobel Moure

Model and test date: 2025 search systems; log conditions and exact routing need checks.

Design and sample: Preprint; LMArena March-April 2025 logs; approximately 14,000 responses, 11 model labels.

What was measured: Visited/credited source attribution, not citation accuracy.

Recorded finding: Studies discrepancies between logged browsing and visible attribution.

Inspected location: Methods; Table 1; limitations.

Limits: Selective logs and inferred source consumption. Does not test claim support or establish copyright/licensing violations.

Qualifications: Repository date/version chronology, full denominators, notices and data access remain to verify. Policy interpretations are the authors' interpretations.

Check coverage: Partial support. Original inspected; dedicated notices and identifier resolution remain unchecked.

Provisional research rating: Limited.

[Original source](#)

Identifier: 10.48550/arXiv.2508.00838.

3. Human trust, checking and error detection

These studies concern people's responses to information, not AI-text detector classification. Perceived trust, opening a source, understanding it, identifying an error and correcting a decision are distinct outcomes. Several studies include beneficial, null and harmful findings under different conditions; those differences are retained.

E016 | Citations and Trust in LLM Generated Responses

Authors: Yifan Ding; Matthew Facciani; Ellen Joyce; Amrit Poudel; Sanmitra Bhattacharya; Balaji Veeramani; Sal Aguinaga; Tim Weninger

Model and test date: ChatGPT4; 13 March 2024.

Design and sample: Randomised citation conditions; Prolific; reported N=303, but group counts sum to 305.

What was measured: Self-reported trust; recorded mouseover proxy, not verified source reading or error correction.

Recorded finding: Citations increased trust; five versus one was not significant. Checking/trust association is correlational.

Inspected location: arXiv v1 Methods; Results; Tables S2-S5.

Limits: Random citations are existing unrelated links, not fabricated publications. Reported denominators need reconciliation; repeated-response analysis needs review.

Qualifications: Publisher metadata confirmed. Full publisher PDF inaccessible through web tool. Preprint title/author order differ; reconcile versions.

Check coverage: Partial support. Original read; dedicated notice checks pending.

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.1609/aaai.v39i22.34550; arXiv:2501.01303.

E017 | “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust

Authors: Sunnie S. Y. Kim; Q. Vera Liao; Mihaela Vorvoreanu; Stephanie Ballard; Jennifer Wortman Vaughan

Model and test date: July 2023 Bing Copilot outputs, wording manipulated; participants September 2023.

Design and sample: Preregistered; US MTurk; 404 retained/656 responses; eight medical questions.

What was measured: Correct final answers; agreement; clicks; separate trust scales.

Recorded finding: Table 1 estimated correctness: control 63.9%, first-person uncertainty 72.8% ($P < .01$).

Inspected location: Sections 3.1-3.5, 4.3-4.5; Tables 1-2; Appendix B.

Limits: Controlled older medical task; no demonstrated effect on UK publishing workflow. Between-condition source usage and trust beliefs showed no significant differences.

Qualifications: Registration existence reported by authors, not independently audited. Exclusions, source-use self-report gap and general-perspective null results retained.

Check coverage: Partial support. Original read; dedicated notice checks pending.

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.1145/3630106.3658941; arXiv:2405.00623.

E018 | Large Language Models Help Humans Verify Truthfulness - Except When They Are Convincingly Wrong

Authors: Chenglei Si; Navita Goyal; Tongshuang Wu; Chen Zhao; Shi Feng; Hal Daumé III; Jordan Boyd-Graber

Model and test date: GPT-3.5-turbo-0613; July-August 2023.

Design and sample: 80 Prolific workers, five groups; 20 difficult claims each.

What was measured: Correct human true/false decisions.

Recorded finding: Wrong-explanation subset: explanation 35%, baseline 49%, retrieval 54%. Overall explanation 74%, retrieval 73%, baseline 59%.

Inspected location: Sections 4.1-4.4 and 5.1-5.3; Figures 2-4; Limitations.

Limits: Small groups; difficult selected claims; old model; statistical and dataset checks pending.

Qualifications: Section 4.4 states $20 \times 16 \times 5 = 1500$; arithmetic gives 1600. Reconcile denominator.

Check coverage: Partial support. Dedicated notices pending.

Provisional research rating: Moderate.

[Original source](#)

Identifier: 10.18653/v1/2024.naacl-long.81.

E019 | Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment

Authors: Sofia Eleni Spatharioti; David M. Rothschild; Daniel G. Goldstein; Jake M. Hofman

Model and test date: GPT-3.5 then GPT-3; field dates unconfirmed.

Design and sample: Randomised SUV decisions: US Mechanical Turk, N90 and N120.

What was measured: Correct vehicle choice.

Recorded finding: Routine choices: model-estimated 92.3% traditional versus 95.3% LLM ($p=.33$); challenging choice 93% versus 47% ($p<.001$). Second experiment: challenging-task accuracy 26% control, 58% dual highlighting, 53% low-only highlighting.

Inspected location: Sections 4-5.1; Figures 5 and 8; footnotes 1-2.

Limits: Selected error-prone task, historical models, constrained interface. Experiments are not directly comparable. An apparent GPT-3.5.1 label in the generated report combined model text with a footnote; the original specifies GPT-3.5.

Check coverage: Partial support. Complete publisher/version and notice checks outstanding

Provisional research rating: Moderate. Controlled task comparison; narrow generalisability.

[Original source](#)

Identifier: arXiv:2307.03744v2.

E020 | Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies

Authors: Sunnie S. Y. Kim; Jennifer Wortman Vaughan; Q. Vera Liao; Tania Lombrozo; Olga Russakovsky

Model and test date: Controlled responses; separate Study 1 used ChatGPT-4o in June 2024.

Design and sample: Study 2: August 2024, US Prolific, 320 recruited, 308 retained; within-subject $2 \times 2 \times 2$ experiment.

What was measured: Objective final-answer correctness; agreement analysed separately.

Recorded finding: Without sources, explanation increased raw accuracy for correct AI answers ($67.2 \rightarrow 78.2\%$) but decreased it for wrong answers ($21.8 \rightarrow 17.2\%$). Sources without explanations reduced wrong-answer agreement ($78.2 \rightarrow 68.2\%$); interaction $\beta = -.83$, $p = .002$.

Inspected location: Sections 4.1.3-4.1.5; Table 1; Sections 4.2.1-4.2.2.

Limits: Sources generally accurate/relevant; no fabricated-citation manipulation. Source clicking is not randomised. Preregistration independently unchecked; final publisher version outstanding.

Check coverage: Partial support. Complete publisher/version and notice checks outstanding

Provisional research rating: Moderate. Controlled behavioural experiment; scope-specific outcomes.

[Original source](#)

Identifier: arXiv:2502.08554v1.

E021 | Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions

Authors: Hyo Jin Do; Rachel Ostrand; Justin D. Weisz; Casey Dugan; Prasanna Sattigeri; Dennis Wei; Keerthiram Murugesan; Werner Geyer

Model and test date: Handcrafted answer and factuality scores; not live-model performance.

Design and sample: Within-subject ratings, 104 company employees across 20 countries; baseline always first.

What was measured: Perceived trust and ease of validation; no objective detection/correction outcome.

Recorded finding: Reference-number annotations raised trust ratings 2.63→3.40 ($p<.001$) while reducing perceived ease of validation 4.29→3.81 ($p=.03$). Highlight gradients showed the same directional pattern.

Inspected location: Sections 3.2-3.3; Table 3; Section 5 limitations.

Limits: One entertainment question, partially incorrect answer, single employer, fixed baseline order. Preprint describes workshop submission, not established peer-reviewed acceptance.

Check coverage: Partial support. Complete publisher/version and notice checks outstanding

Provisional research rating: Limited. Direct perception measures; no demonstrated improvement in actual error detection.

[Original source](#)

Identifier: arXiv:2405.20434v1.

E022 | To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models

Authors: Jessica Y. Bo; Sophia Wan; Ashton Anderson

Model and test date: GPT-4o advice, adjusted and sampled; field dates unconfirmed.

Design and sample: Randomised four-condition experiment; 400 US Prolific recruits, 397 retained (group totals).

What was measured: Appropriate reliance ratio (ARR), not overall accuracy.

Recorded finding: LSAT: disclaimer ARR1.62 versus control1.17, permutation $p=.04$; other intervention ARR comparisons non-significant. Estimation ARR comparisons non-significant. Overall task-performance differences between interventions and control non-significant.

Inspected location: Sections 3.2-3.3, 4.1-4.3; Table 1.

Limits: Good/bad advice deliberately balanced, fixed stimuli, no external checking or follow-ups. Preregistration and publisher checks outstanding.

Check coverage: Partial support.

Provisional research rating: Moderate.

[Original source](#)

Identifier: arXiv:2412.15584v3; first preprint December 2024, revised October 2025.

E023 | Critical or Compliant? The Double-Edged Sword of Reasoning in Chain-of-Thought Explanations

Authors: Eunkyu Park; Wesley Hanwen Deng; Vasudha Varadarajan; Mingxi Yan; Gunhee Kim; Maarten Sap; Motahhare Eslami

Model and test date: Constructed reasoning perturbations and tone edits; field date unconfirmed.

Design and sample: Reported 800 US AMT workers; single multimodal moral-judgement stimulus per participant.

What was measured: Binary error flag, agreement and reported trust; no consequential downstream behaviour.

Recorded finding: Confident tone reduced error-detection log-odds ($\beta=-.49$); agreement/trust tone effects null. Table 4: hedged detection50.4%, confident40.5%.

Inspected location: Section 4 Q2, Tables 2 and 4; Appendix A.5 and A.8.4.

Limits: Post-exclusion N unclear. Table3 omission miscalibration27.5% differs from Table9 verdict strata42%/35%; The inconsistent statistic is not adopted here. Conference acceptance self-reported; final proceedings unverified.

Check coverage: Partial support.

Provisional research rating: Limited.

[Original source](#)

Identifier: arXiv:2511.12001v3; revised 28 September 2026, first preprint November 2025.

4. Craft, voice, cultural expression and skills

This section contains the initial craft review and the extension's selected original-source checks. Professional accounts, controlled short tasks and educational experiments contribute different kinds of evidence. No data or code were rerun. Historical tools are identified, and qualified leads remain separate from adopted findings.

C01 | Ideas for microfiction: benefits and anchoring

Anil R. Doshi and Oliver P. Hauser, *Generative AI enhances individual creativity but reduces the collective diversity of novel content*, Science Advances 10, eadn5290, 12 July 2024. [Publication DOI](#); [author manuscript, v3](#); [inspected journal PDF copy](#).

A randomised study analysed 293 UK participants' eight-sentence stories and ratings from 600 evaluators. Conditions offered no AI idea, one GPT-4 idea, or up to five. Access to ideas improved rated novelty and usefulness, particularly for participants with lower baseline divergent-association scores. Assisted stories were also more similar within conditions.

Support inspected: journal pp. 1-4, Figures 1-4; limitations p. 6; recruitment and task pp. 7-8. Author-manuscript methods corroborate the sample and task. The journal site was unavailable through the web tool, so a separately hosted journal PDF was used and its title, authors and DOI cross-checked.

Limits: participants were not selected as professional writers. A brief, fixed-prompt task does not establish novel quality, sustained skill gains or commercial success. Creativity ratings are not copyright originality tests. Published supplements and raw data still need full audit.

C02 | Co-writing diversity depends on the model

Vishakh Padmakumar and He He, *Does Writing with Language Models Reduce Content Diversity?*, ICLR 2024. [Conference paper](#).

Thirty-eight Upwork writers with writing or copyediting experience produced 300 argumentative essays across ten topics, with approximately 300 words requested per essay. Each session included solo writing, GPT-3 davinci and InstructGPT text-davinci-003, with randomised condition order and topic assignment.

InstructGPT co-writing increased similarity and reduced lexical/content diversity. The corresponding similarity increase was not statistically significant for GPT-3. Analysis attributed much of the reduction to model-contributed material; human-contributed portions did not show the same measured change.

Support inspected: sections 2-3, pp. 2-4; sections 4-6, pp. 5-8, Figure 4; Appendix D, p. 28.

Limits: historical models, essay tasks and a continuation interface. Some content measures use model-generated summaries and attribution rules. No direct test of modern conversational editing, individual literary voice, long-term learning or novels. A non-significant result is not proof of no possible effect.

C03 | Professionals' experience with Dramatron

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman and Richard Evans, *Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals*. [Inspected author preprint, 29 September 2022](#); [CHI 2023 publication DOI](#).

Fifteen theatre and film professionals used the Chinchilla-based Dramatron system in approximately two-hour co-writing sessions and interviews. Participants described inspiration and alternative ideas, alongside repetition, stereotypes, consistency problems, weak subtext and character motivation.

Support inspected: preprint sections 3.1, 3.4 and 4 for model, interaction and participants; sections 5.2-5.5 for interview themes.

Limits: a small qualitative, developer-led study with participants already experienced in AI writing tools. Their observations are not measured improvements over a no-AI control or representative estimates for authors. The inspected text is the 2022 preprint; the published CHI version still requires comparison because the publisher page was inaccessible. Historical limitations should not be asserted as universal current capabilities. Participants' sense of ownership does not determine legal ownership.

C04 | Current linguistic diversity study: abstract-only lead

Zhivar Sourati and colleagues, *The shrinking landscape of linguistic diversity in the age of large language models*, *Nature Human Behaviour*, published 24 August 2026. [Original journal page](#).

The abstract reports reduced stylistic variation across multiple datasets and LLM rewriting conditions. Metadata, abstract, figure captions and data/code availability were inspected. The methods and complete results were behind subscription access. **Headline percentages and conclusions about novels are not adopted here: full methods, attribution, model and comparison-group checks remain incomplete.**

C05 | Contrary lead: stylistic adaptation

Vivan Doshi and Mengyuan Li, *Writing in Symbiosis: Mapping Human Creative Agency in the AI Era*. [Inspected preprint](#).

Sections 2-4 describe observational arXiv/Discord analysis and different stylistic trajectories. The limitations explicitly identify non-causal analysis and constructs not validated through direct participant studies. Peer-reviewed publication status was not confirmed. **Statistical similarity and calendar periods do not establish an individual's actual AI use or intention.** This remains a contrary discovery lead, rather than evidence that assisted authors preserve voice.

A006-S01: Professional writers over eight weeks

Daphne Ippolito, Ann Yuan, Andy Coenen and Sehmon Burnam, [Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers](#), author manuscript v1, 9 November 2022. Sections 3, 4.1, 6, 7 and 9 inspected.

Thirteen published writers used LaMDA-based Wordcraft across eight weeks to produce 1,000-1,500-word stories; journals and interviews supplied qualitative evidence. Useful brainstorming and experimental workflows coexisted with voice, repetition and understanding problems.

Selected-text support within this small historical study. Google developer-led, selected participants, no randomised comparison or retained-skill test. Eight weeks is sustained engagement, not a book-length trial. Tool/interface and model effects cannot be separated. Published-version reconciliation and notices remain pending.

A006-S02: Writers embracing and resisting AI

Rama Adithya Varanasi, Batia Mishan Wiesenfeld and Oded Nov, [AI Rivalry as a Craft](#), author manuscript v1, 12 March 2025, displaying CHI 2025 DOI 10.1145/3706598.3714035. Sections 3, 4 and 5.4 inspected.

Twenty-five US writing professionals interviewed March-June 2024; varied roles and at least twelve months' prior AI awareness. Some described useful delegation; others deliberately protected human craft and identity. AI management itself required work.

Selected-text support for qualitative accounts, not measured productivity or learning. Selected participants had access and autonomy; accounts are not representative industry prevalence or causal outcomes. Exact model builds were not experimentally controlled. Publisher-version comparison, supplements and notices remain pending.

A006-S03: Queer and disabled fiction writers

Carolyn Lamb, Daniel G. Brown and Maura R. Grossman, [Precarity and Solidarity](#), IJCAI 2025, DOI 10.24963/ijcai.2025/1124. Methods pp.10117-10118, results and limitations p.10123 inspected.

May 2024 snowball survey: 137 complete/partial responses; eight excluded, leaving 129 (54 queer-disabled writers and 75 others). Qualitative analysis focused on 16 queer-disabled respondents. Respondents expressed concerns about precarity and AI's impact.

Selected-text support for selected experiences. Not a test of assistive benefit, skill loss or actual market displacement. Recruitment overrepresented queer/disabled writers; nationality not collected, predominantly white sample. Authors' positionality disclosed; sensitive raw qualitative data withheld. No claim about every disabled author is justified.

A006-S04: Creative-writer boundaries

Alicia Guo, Shreya Sathyanarayanan, Leijie Wang, Jeffrey Heer and Amy Zhang, [From Pen to Prompt](#), author v4, 10 June 2025; C&C DOI 10.1145/3698061.3726910. Sections 3, 4.1 and 6 inspected.

Eighteen existing AI users interviewed November 2023-July 2024; observed writing averaged 34.5 minutes. Participants described task boundaries, voice, ownership and enjoyment.

Selected-text support for accounts and observed workflows, not causal preservation of craft.

Recruitment favoured users; no non-user, book-quality or unaided-skill comparison. Ownership here is a reported feeling, not copyright. Residence counts total 19 despite stated sample 18; no location distribution adopted. Publisher comparison and notices pending.

A006-S05: Performance after assistance removal

Sarah Shi Hui Wong and Sophia Xuefei Qiu, [Think First, ChatGPT Later](#), published 15 April 2026. Method, Results and Limitations inspected.

197 Singapore students recruited, 196 analysed: randomised human-only 64, unrestricted AI 67, guided AI 65. Reported ChatGPT-4; two 12-minute product tasks, blind expert scoring. Guided users performed better on the subsequent unaided task; unrestricted users did not significantly outperform controls then.

Selected-text support for this short experiment; INDIRECT for authors. Guidance combined independent ideation, AI development and independent refinement. No equally guided human-only condition or delayed follow-up. This does not isolate AI's learning benefit or prove permanent deskilling.

A006-S06: Cultural expression

Dhruv Agarwal, Mor Naaman and Aditya Vashistha, [AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances](#), author v3, 12 March 2025; CHI DOI 10.1145/3706598.3713564. Sections 3, 4.3 and 5.4 inspected.

118 Prolific participants, 60 India and 58 US, randomised to GPT-4o autocomplete or no AI; four English tasks, minimum 50 words. Authors report greater cross-group similarity and movement towards US patterns.

Selected-text support for tested group comparisons. Not books, all cultures or individual voice loss. Pairwise text comparisons exceed participant counts; dependence-aware statistical review remains pending. Country proxies, English fluency and selected recruitment limit transfer. Funding includes Infosys and Microsoft. Published-version comparison and notices pending.

A006-S07: Whole-novel lead

The linked [AutoFiction workshop PDF](#) was inaccessible to the web tool. [Author project documentation](#) confirms an evolving automated novel-generation pipeline with limited human involvement, and [platform information](#) confirms the research project.

Discovery lead only. The generated report's book and reader counts were not independently confirmed in the original paper and are not adopted here. This automated pipeline is not evidence of sustained collaborative performance by human authors.

5. Detectors and publishing policies

The evidence here concerns specific detector tests and two organisations' stated policies. It establishes neither industry-wide enforcement practice nor the history of any individual manuscript. Recommendations for a fair assessment are editorial proposals; they do not create a legal entitlement to publication.

A005-S01: Current detector evaluation

[Jaeho Lee, Epoch AI, 15 July 2026](#). Original webpage inspected: error-rate table, Data, Analysis, and Assumptions and limitations.

Selected-text support for the reported test counts: 495 human passages of about 500 words from 99 authors; 297 generations per prompting condition. Human false positives: Pangram 0/495, GPTZero 0/495, [Originality.ai](#) 19/495. Basic-prompt misses: 0/297, 2/297, 1/297; style-imitation misses: 30/297, 32/297, 53/297. The recorded comparison used specified detector builds and generator models.

Moderate evidence within this test. Independent online analysis; peer review not established. Human passages cluster by author; texts are short and do not establish whole-book or mixed-writing performance. "Mixed" counted as a miss; treating it as detection changes two rates. Zero observed false positives does not establish zero population risk. Webpage support is checked; repository provenance and reported intervals have not been independently reproduced.

This is a recent independent test; a claim that it is the strongest test available has not been established.

A005-S02: Historical non-native writing experiment

[Liang, Yuksekgonul, Mao, Wu and Zou, author manuscript v3, 10 July 2023](#). Inspected Results and Figure 1/2 discussion.

Selected-text support: seven detectors, 91 TOEFL essays and 88 US eighth-grade essays; mean TOEFL false-positive rate **61.22%**; 18/91 flagged by all seven and 89/91 by at least one. The latter is not one detector's error rate. The manuscript reports simplification changing mean misclassification from 5.19% to 56.65%, and generated-essay detection changing from 100% to 13% after its self-edit intervention.

Moderate historical support for this specific finding. Student groups differ beyond language background. This does not establish today's rates, all-language bias or performance on novels. Inspected the original author manuscript linked to its published version, not a full journal/supplement audit. Later corrections and detector-version reconciliation remain pending.

A005-S03: RAID citation mismatch

[Correct original publication record: Dugan et al., RAID, ACL August 2024](#), DOI 10.18653/v1/2024.acl-long.674, pages 12463-12492.

Source identification: the original discovery references confused RAID with the Weber-Wulff paper. The corrected ACL publication record is used here.

The ACL abstract supports the benchmark description (over six million generations; 11 models, eight domains, 11 attacks, four decoding strategies; eight open and four closed detectors) and its broad robustness finding. **Partial support only** until methods and result tables are inspected. Additional Perkins detector-study tables were not independently checked.

A005-S04: Agency submission policy

[Greene & Heaton official submission guidelines](#), inspected opening two requirements paragraphs and “What happens next”.

Selected-text support for this specific finding: the live page rejects AI-originated, written or edited submissions, including the cover letter, synopsis, proposal and manuscript. It describes perceived effects of editing on writing, but supplies no named detector, test threshold or case-level decision record.

Strong for this organisation’s stated rule; no industry prevalence finding. Undated live wording checked on 6 October 2026; earlier policy versions and enforcement not established. A submission restriction does not itself prove false accusations or detector reliance.

A005-S05: Publisher submission policy

[The Book Dragon Publishing Press official page](#), “AI and Generated Content”, all four paragraphs.

Selected-text support for this stated policy: the page requires disclosure, distinguishes assistance from generation, welcomes human-led assistance subject to individual assessment, and excludes work it considers wholly or heavily generated. It also states rejection or contract termination for undisclosed use. That is policy wording, not a determination that termination is legally valid in every case.

Strong for stated wording only. The effective date of the AI wording has not been independently established. No named detector or individual case record was inspected. Other publisher policies remain outside the completed checking.

Base rates and interpretation

A score is not automatically a probability that a particular author used AI. A detector’s sensitivity, false-positive rate and the prevalence of AI text in the assessed population all affect interpretation. The discovery report’s hypothetical positive-predictive-value examples are arithmetic illustrations, not measured publisher prevalence or calibrated author scores.

Writing quality, factual accuracy, policy compliance, copyright authorship and statistical text classification need separate assessments. Draft and revision records can supply context, but are not guaranteed legal protection. This companion adopts no claim about named-author misconduct or publisher wrongdoing.

6. Rights, confidentiality and legal-source boundaries

These are dated source notes, not a complete legal opinion. Statutes, court judgments, regulator guidance, consultations, proposals and law-firm commentary have different authority. Access limitations and incomplete current-text or later-treatment checks are retained. Application to a book, contract or dispute needs the relevant facts and qualified advice.

L001: Lewis Silkin briefing

[IP and AI: when innovation becomes infringement](#), Oliver Fairhurst and Adrian Aronsson-Storrier. Displayed publication date: 2026-10-01; URL date differs.

Status: law-firm commentary and practical advice, not binding law. Location: final practical-advice section. Supports meaningful human authorship, process documentation, output clearance and review of indemnity/insurance. It does not guarantee copyright or non-infringement.

L002: Historical official explanation

[Artificial Intelligence and Intellectual Property: copyright and patents](#), Intellectual Property Office. Updated 2022-06-28; consultation began in 2021.

Status: historical consultation document, not new legislation. Location: Copyright → Existing copyright protection. Explains protection for AI-assisted work expressing original human creativity. This passage concerns human-assisted work; consultation options and computer-generated-work provisions are separate matters.

L003: More recent government policy

[Report on Copyright and Artificial Intelligence](#), DSIT, DCMS and IPO, 2026-03-18.

Status: policy report presented under the Data (Use and Access) Act 2025, not an amendment to copyright law. Section I distinguishes existing computer-generated-work protection from proposals to remove it, while retaining protection for AI-assisted human work. A proposal must not be labelled enacted law. Later legislative developments remain to check.

L004: CMA unfair commercial practices

[Official guidance](#), CMA207, updated November 2025. Overview and core concepts inspected. The CMA says DMCC unfair-trading provisions apply to commercial practices from April 2025. Misleading actions generally require the specified conduct and likely effect on the average consumer's transactional decision. Some banned practices are prohibited regardless of that effect. This is regulator guidance, not the statutory text. Its application to a publishing scenario needs the facts and relevant legal elements.

Lo05: Ayinde and Al-Haroun

[Original approved judgment](#), [2025] EWHC 1383 (Admin), England and Wales Divisional Court. Selected opening paragraphs inspected: especially 7-9 on checking AI-assisted legal research and lawyers' supervisory responsibilities. This is a court judgment concerning legal proceedings and legal professionals. It does not establish a general publishing verification duty. Complete disposition, later treatment and application to different professions remain to check.

Lo06: ICO accuracy guidance

[Original ICO chapter](#). Displayed chapter update: March 2023. Opening notice and accuracy-versus-statistical-accuracy section inspected. The ICO distinguishes personal-data accuracy from model performance and does not require every AI prediction to be correct. It discusses recording inference status and provenance. The page expressly says guidance is under review following the Data (Use and Access) Act. This is regulator guidance; current legislation, journalistic/literary exemptions and application to publishing require separate checks.

Lo07: 2026 AI code regulations

[Official instrument PDF](#), SI 2026/425. Regulations 1-3 and the explanatory note inspected. Made 16 April, laid 21 April; commencement stated as 12 May 2026. Extends throughout the UK. Regulation 2 requires the Commissioner to prepare good-practice guidance on personal-data processing for AI development/use and automated decision-making, including children's data. Relevant legislation excludes DPA Part 4. This supports the specific claim about the ICO's code-preparation duty. It does not establish a completed code or a universal publisher-checking obligation. Later amendments and the actual code's preparation/publication status remain unchecked.

What remains unresolved

Current amended statutory text, commencement, territorial scope and later judicial treatment have not been comprehensively audited. Several official statutory pages were inaccessible during the research workflow. The discovery report's consumer-redress commencement claim was not independently confirmed. No universal author, editor or publisher verification duty is inferred from this research.

Training-data rights, the author's human creative contribution, output infringement, privacy, confidentiality and contractual disclosure should be considered separately. Keeping records does not itself create copyright, establish non-infringement or prove compliance with every agreement.

7. Proposition map and boundaries

The thirteen propositions below connect the narrative to its supporting source records. They are bounded research propositions and editorial recommendations, rather than final human-approved conclusions. P-numbers identify propositions; E-numbers identify evidence-register records; C and A references identify source-check notes in this companion.

Po1 | AI ideas can improve ratings of brief stories while outputs become more alike.

Evidence: Doshi and Hauser, C01; 2024, GPT-4; 293 writers, 600 evaluators.

Location and coverage: Main results, Figs 1-4; methods pp.7-8. Selected original text checked.

Boundary: Short fixed task; not complete books or legal originality.

Po2 | Diversity outcomes depend on the tested model and assistance.

Evidence: Padmakumar and He, C02; ICLR 2024; 38 writers, 300 essays.

Location and coverage: Sections 2-6 and Fig.4. Selected methods/results checked.

Boundary: Historical continuation interface; model-derived metrics; no long-term learning.

Po3 | Professional writers can find useful workflows and voice problems.

Evidence: Ippolito et al., A006-S01; author v1 2022.

Location and coverage: Sections 3-4, 6-7, 9. Selected original text checked.

Boundary: 13 selected writers, LaMDA, eight weeks; qualitative and developer-led.

Po4 | Identifiable citations can still misrepresent evidence.

Evidence: Linardon et al., E007; December 2025 test, reported ChatGPT-5 thinking model.

Location and coverage: Methods and Results. Selected claim-support audit checked; The generated discovery link was corrected.

Boundary: 12 mental-health reviews, 333 citation-claims; conservative passage scoring; not general author error rate.

P05 | Helpful overall AI explanations can coexist with harm when explanations are wrong.

Evidence: Si et al., E018; NAACL 2024; GPT-3.5-turbo-0613 test in 2023.

Location and coverage: Sections 4-5; Figs 2-4. Conditional comparison supported.

Boundary: Difficult selected claims, small groups; sample arithmetic remains unresolved.

P06 | Detector performance differs by condition; scores do not reconstruct manuscript history.

Evidence: Lee, A005-S01; Epoch AI, July 2026.

Location and coverage: Error-rate table, Data, Analysis, limitations. Webpage counts checked.

Boundary: 495 human passages clustered by 99 authors; 297 generations per condition; raw-data replication pending.

P07 | Two UK submission policies describe different approaches to AI.

Evidence: Greene and Heaton and Book Dragon, A005-S04-S05.

Location and coverage: Official submission wording checked on 6 October 2026.

Boundary: Stated rules only; not enforcement, industry prevalence or legal validity of termination.

P08 | Human creative contribution, documentation and IP clearance deserve separate attention.

Evidence: L001-L003, official explanation/policy and Lewis Silkin commentary.

Location and coverage: Existing copyright protection; 2026 Section I; final practical advice.

Boundary: Source passages checked; amended statutes, later treatment and case-specific interpretation pending.

P09 | Authors' views of AI include craft, autonomy and livelihood concerns.

Evidence: Varanasi et al., A006-S02; Lamb et al., A006-S03.

Location and coverage: Selected methods/findings/limitations checked.

Boundary: Qualitative selected populations; reported experience is not measured market displacement.

P10 | The seven-step workflow is proposed editorial practice.

Evidence: Narrative section 7 and the selected evidence in this companion.

Location and coverage: Reasoned synthesis, not an independently tested intervention.

Boundary: No guaranteed accuracy, skill retention, copyright or contract compliance.

P11 | Experienced users describe boundaries around assistance.

Evidence: Guo et al., A006-S04.

Location and coverage: Sections 3, 4.1, 6; selected original passages checked.

Boundary: Selected qualitative accounts; no causal skill or rights finding.

P12 | Guided AI use improved immediate unaided task performance.

Evidence: Wong and Qiu, A006-S05.

Location and coverage: Method, Results, Limitations checked.

Boundary: Student product design; no delayed or matched-guidance control.

P13 | Autocomplete can affect cultural expression in short tasks.

Evidence: Agarwal et al., A006-S06.

Location and coverage: Sections 3, 4.3, 5.4 checked.

Boundary: Selected India/US English tasks; statistical review pending.

Conclusions this research does not establish

This research does not establish a general accuracy rate or universal improvement rate for today's AI systems; permanent loss or improvement of professional authors' skills; universal loss of voice; universal detector failure; or conclusive reconstruction of a book's provenance from a score.

It also does not establish industry-wide prevalence of detector sanctions, wrongdoing by a named publisher, automatic copyright or contractual compliance through record keeping, or a universal legal duty to verify every factual statement.

8. Unresolved evidence and future research

These gaps belong in the public account of the research. Their inclusion does not imply that no evidence exists; it means this project has not completed the checking needed to adopt a stronger conclusion.

Sustained books and professional practice

Controlled evidence on complete novels, substantial non-fiction, modern developmental editing and long-term professional practice remains limited in the selected material. A short exercise or automated novel pipeline is not interchangeable with sustained human authorship.

Retained learning

A better assisted output does not establish better independent performance later. Immediate unaided task results need to be separated from delayed retention, transfer and dependence. Guided AI use also needs an equally guided human-only comparison.

Accessibility and language

Selected accounts from disabled, queer and multilingual writers do not establish universal assistive benefit or harm. Appropriate recruitment, consent, task design and outcomes are needed.

Version and statistical checks

Author manuscripts need comparison with final publications. Denominator inconsistencies, dependent text comparisons, summary/table disagreements, automated scoring calibration and null outcomes require explicit attention.

Policy and disputed decisions

Two organisations' stated rules are not an industry survey. Stronger claims about sanctions need dated policies, primary decision records, the author's response and the actual contractual terms.

Independent review

Priority original methods and findings, supplements, correction notices, identifiers and relevant legal authorities still need human review. No exhaustive correction or retraction search or raw-data replication is claimed.

Additional discovery leads not adopted as findings

[2026 EFL creative-writing study](#)

Discovery lead only. Full methods, exact support, versions and applicability have not been established in this companion; no numerical finding is adopted here.

[2025 retained-creativity study](#)

Discovery lead only. Full methods, exact support, versions and applicability have not been established in this companion; no numerical finding is adopted here.

[Chinese-writing study](#)

Discovery lead only. Full methods, exact support, versions and applicability have not been established in this companion; no numerical finding is adopted here.

[Homogenisation systematic review and meta-analysis](#)

Discovery lead only. Full methods, exact support, versions and applicability have not been established in this companion; no numerical finding is adopted here.

[Ethics and the Use of Generative AI in Professional Editing](#)

Discovery lead only. Full methods, exact support, versions and applicability have not been established in this companion; no numerical finding is adopted here.

[The HaLLMark Effect: writing, provenance](#)

Discovery lead only. Full methods, exact support, versions and applicability have not been established in this companion; no numerical finding is adopted here.

9. References and original-source links

These links preserve the original-source trail used in the dossier. A listed link is not a claim of complete verification. Preprints, specialist benchmarks, organisational policies and legal commentary retain the qualifications in their corresponding record. Publisher or author-hosted sources are preferred to generated citation links.

E001 | Fabrication and errors in the bibliographic citations generated by ChatGPT

William H Walters; Esther Isabelle Wilder. 2023

<https://pmc.ncbi.nlm.nih.gov/articles/PMC10484980/>

E002 | Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis

Mikaël Chelli and colleagues. 2024

<https://pmc.ncbi.nlm.nih.gov/articles/PMC11153973/>

E003 | Accuracy of Chatbots in Citing Journal Articles

Anjun Chen; Drake O Chen. 2023

<https://pmc.ncbi.nlm.nih.gov/articles/PMC10410472/>

E004 | Hallucination Rate of Peer-Reviewed Citations Generated by Large Language Models in Neurocritical Care

Ali Seifi; Ali Seyfi. 2026

<https://pmc.ncbi.nlm.nih.gov/articles/PMC13506236/>

E005 | Synthesizing scientific literature with retrieval-augmented language models

Akari Asai and colleagues. 2026

<https://www.nature.com/articles/s41586-025-10072-4>

E006 | Citation Constraints and Reference Hallucinations in Large Language Models

Kimberly Davis; Qusay Mahmoud. 2026

<https://proceedings.mlr.press/v318/davis26a.html>

E007 | Fluency Without Fidelity: Errors in Citation-Attributed Claims in Large Language Model-Generated Literature Reviews in Mental Health

Jake Linardon; Mariel Messer; Cleo Anderson; Olivia Marie Soliman; Claudia Liu; Joseph Firth; John Torous. 2026

<https://link.springer.com/article/10.1007/s41347-026-00630-6>

E008 | Large Language Models Require Curated Context for Reliable Political Fact-Checking; Even with Reasoning and Web Search

Matthew R. DeVerna; Kai-Cheng Yang; Harry Yaojun Yan; Filippo Menczer. Publication/version details incomplete.

https://www.matthewdeverna.com/docs/publications/2026_DeVerna_LLMFCing/2025_DeVerna_LLMragFCing.pdf

E009 | This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance

Courtnei Byun; Piper Vasicek; Kevin Seppi. 2024

<https://aclanthology.org/2024.hcinlp-1.3/>

E010 | Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties

Yasin Celal Güneş; Turay Cesur; Eren Çamur. 2025

<https://www.dirjournal.org/articles/evaluating-the-reference-accuracy-of-large-language-models-in-radiology-a-comparative-study-across-subspecialties/doi/dir.2025.253101>

E011 | Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation

Sirui Xia; Xintao Wang; Jiaqing Liang; Yifei Zhang; Weikang Zhou; Jiaji Deng; Fei Yu; Yanghua Xiao. 2025

<https://aclanthology.org/2025.findings-naacl.55/>

E012 | Influence of Topic Familiarity and Prompt Specificity on Citation Fabrication in Mental Health Research Using Large Language Models: Experimental Study

Jake Linardon; Hannah K Jarman; Zoe McClure; Cleo Anderson; Claudia Liu; Mariel Messer. 2025

<https://mental.jmir.org/2025/1/e80371/>

E013 | CiteFix: Enhancing RAG Accuracy Through Post-Processing Citation Correction

Harsh Maheshwari; Srikanth Tenneti; Alwarappan Nakkiran. 2025

<https://aclanthology.org/2025.acl-industry.23/>

E014 | Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents

Hailey Onweller; Elias Lumer; Austin Huber; Pia Ramchandani; Vamse Kumar Subbiah; Corey Feld. 2026

<https://arxiv.org/abs/2605.06635>

E015 | The Attribution Crisis in LLM Search Results

Ilan Strauss; Jangho Yang; Tim O'Reilly; Sruly Rosenblat; Isobel Moure. 2025

<https://arxiv.org/abs/2508.00838>

E016 | Citations and Trust in LLM Generated Responses

Yifan Ding; Matthew Facciani; Ellen Joyce; Amrit Poudel; Sanmitra Bhattacharya; Balaji Veeramani; Sal Aguinaga; Tim Weninger. 2025

<https://ojs.aaai.org/index.php/AAAI/article/view/34550>

E017 | “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust

Sunnie S. Y. Kim; Q. Vera Liao; Mihaela Vorvoreanu; Stephanie Ballard; Jennifer Wortman Vaughan. 2024

<https://arxiv.org/html/2405.00623v1>

E018 | Large Language Models Help Humans Verify Truthfulness - Except When They Are Convincingly Wrong

Chenglei Si; Navita Goyal; Tongshuang Wu; Chen Zhao; Shi Feng; Hal Daumé III; Jordan Boyd-Graber. 2024

<https://aclanthology.org/2024.naacl-long.81/>

E019 | Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment

Sofia Eleni Spatharioti; David M. Rothschild; Daniel G. Goldstein; Jake M. Hofman. 2023

<https://arxiv.org/pdf/2307.03744>

E020 | Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies

Sunnie S. Y. Kim; Jennifer Wortman Vaughan; Q. Vera Liao; Tania Lombrozo; Olga Russakovsky. 2025

<https://arxiv.org/html/2502.08554v1>

E021 | Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions

Hyo Jin Do; Rachel Ostrand; Justin D. Weisz; Casey Dugan; Prasanna Sattigeri; Dennis Wei; Keerthiram Murugesan; Werner Geyer. 2024

<https://arxiv.org/html/2405.20434v1>

E022 | To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models

Jessica Y. Bo; Sophia Wan; Ashton Anderson. 2025

<https://arxiv.org/html/2412.15584v3>

E023 | Critical or Compliant? The Double-Edged Sword of Reasoning in Chain-of-Thought Explanations

Eunkyu Park; Wesley Hanwen Deng; Vasudha Varadarajan; Mingxi Yan; Gunhee Kim; Maarten Sap; Motahhare Eslami. 2026

<https://arxiv.org/html/2511.12001v3>

Publication DOI

<https://doi.org/10.1126/sciadv.adn5290>

author manuscript, v3

<https://arxiv.org/abs/2312.00506v3>

inspected journal PDF copy

<https://www.rivista.ai/wp-content/uploads/2025/05/sciadv.adn5290.pdf>

Conference paper

https://proceedings.iclr.cc/paper_files/paper/2024/file/02dec8877fb7c6aa9a79f81661baca7c-Paper-Conference.pdf

Inspected author preprint, 29 September 2022

<https://arxiv.org/html/2209.14958v1>

CHI 2023 publication DOI

<https://doi.org/10.1145/3544548.3581225>

Original journal page

<https://www.nature.com/articles/s41562-026-02550-0>

Inspected preprint

<https://arxiv.org/html/2512.13697v1>

Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers

<https://arxiv.org/html/2211.05030v1>

AI Rivalry as a Craft

<https://arxiv.org/html/2503.09901v1>

Precarity and Solidarity

<https://www.ijcai.org/proceedings/2025/1124.pdf>

From Pen to Prompt

<https://arxiv.org/html/2411.03137v4>

Think First, ChatGPT Later

<https://link.springer.com/article/10.1007/s10648-026-10118-7>

AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances

<https://arxiv.org/html/2409.11360v3>

AutoFiction workshop PDF

https://genaicreativity.org/icml2026/files/22/22_paper.pdf

Author project documentation

<https://github.com/AutoFiction-AI/autofiction>

platform information

<https://www.autofiction.ai/about>

Jaeho Lee, Epoch AI, 15 July 2026

<https://epoch.ai/data-insights/ai-detectors-false-negatives>

Liang, Yuksekgonul, Mao, Wu and Zou, author manuscript v3, 10 July 2023

<https://arxiv.org/html/2304.02819v3>

Correct original publication record: Dugan et al., RAID, ACL August 2024

<https://aclanthology.org/2024.acl-long.674/>

Greene & Heaton official submission guidelines

<https://greeneheaton.co.uk/submissions>

The Book Dragon Publishing Press official page

<https://thebookdragon.co.uk/the-book-dragon-publishing-press/>

IP and AI: when innovation becomes infringement

<https://www.lewissilkin.com/insights/2026/10/02/ip-and-ai-when-innovation-becomes-infringement>

Artificial Intelligence and Intellectual Property: copyright and patents

<https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents>

Report on Copyright and Artificial Intelligence

<https://www.gov.uk/government/publications/report-and-impact-assessment-on-copyright-and-artificial-intelligence/report-on-copyright-and-artificial-intelligence>

Official guidance

<https://www.gov.uk/government/publications/unfair-commercial-practices-cma207/unfair-commercial-practices>

Original approved judgment

<https://www.judiciary.uk/wp-content/uploads/2025/06/Ayinde-v-London-Borough-of-Haringey-and-Al-Haroun-v-Qatar-National-Bank.pdf>

Original ICO chapter

<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/what-do-we-need-to-know-about-accuracy-and-statistical-accuracy/>

Official instrument PDF

https://www.legislation.gov.uk/uksi/2026/425/pdfs/uksi_20260425_en.pdf