AI and entry-level knowledge work: Evidence from research assistants*

Johan Fourie†

Abstract

Studies of AI in science observe isolated tasks or the authors on a byline. They do not observe the research assistants whose data cleaning, coding, literature search, and replication work language models now perform. Economics working papers record this junior labour in their acknowledgements, next to the full text. Using 24 working-paper series through October 2026, I compare papers by the same economist that do and do not disclose AI use, alongside a detector of AI-like prose. Disclosing papers credit 0.548 more named assistants than other papers by the same economist, two-thirds of the mean, and run longer, mostly in their appendices. Disclosures of operational use, such as coding and data work, carry the whole assistant difference; writing-only disclosures and AI-like prose show no consistent association with assistants or pages. The paper brings acknowledged junior labour and selective team disclosure into the study of scientific production. Within an economist's portfolio, operational AI accompanies wider, assistant-supported projects.

Keywords: artificial intelligence; research assistants; entry-level work; apprenticeship; knowledge work; scientific teams; voluntary disclosure

JEL codes: J23; J24; O33; A11; D82

*I thank Lukas Althoff, Joshua Gans, Alex Imas, Benjamin Jones, and Preston McAfee for feedback on the paper. I also acknowledge feedback provided by refine.ink. Anthropic's Claude Fable and Google's Gemini 4 Argon helped with drafting and editing the text, and several models, including OpenAI's GPT-6 Astra, were used for coding and analysis, as Appendix A describes. All data come from publicly accessible sources. Cite this paper as: Fourie, Johan. 2026. "AI and entry-level knowledge work: Evidence from research assistants." Working Paper, Department of Economics, Stellenbosch University.

†Department of Economics, Stellenbosch University. Email: johanf@sun.ac.za.

1 Introduction

Before an economics paper reaches a working-paper series, someone cleans the raw files, writes the estimation code, searches the literature, and assembles the appendices. That work has long belonged to the research assistant. Large language models now do much of it in minutes (Korinek, 2023, 2025). Payroll and résumé studies find lower entry-level hiring in occupations exposed to AI (Brynjolfsson, Chandar and Chen, 2026; Hosseini and Lichtinger, 2026), and Garicano and Rayo (2025) show how automation can erode the apprenticeship that pays for a junior's training. Studies of AI in science, however, examine isolated tasks or the authors on a byline (Noy and Zhang, 2023; Kusumegi et al., 2025; Wang et al., 2026). None observes the research assistants credited behind a paper.

I ask whether generative AI substitutes for or complements the research assistant in economics working papers. Working papers record assistance in their acknowledgements, next to the byline, and preserve the full text. Across 56,585 records in 24 series through October 2026, 23 institutional series and the economics arXiv, I link assistance credits to two AI indicators: an explicit statement that the authors used AI, and a writing flag from a detector of AI-like economics prose (Goldsmith-Pinkham, 2026; Jabarian and Imas, 2025). The main regressions compare post-2022 papers written by the same recurring researcher, with researcher, source-quarter, and field fixed effects. Comparing papers within one researcher's portfolio removes persistent differences in funding and in the habit of thanking people. The credits measure contributions to a paper, not hours or employment.

Within an economist's portfolio, papers that disclose AI use credit more research assistants and run longer. Disclosing papers credit 0.548 more named assistants, two-thirds of the mean of 0.83 on papers without a disclosure, and they run about a tenth longer. Both differences survive the adjustment for the registered family of twenty tests. Disclosing papers are also more likely to credit any assistant and carry slightly more authors, but these differences do not survive the adjustment. The disclosure variation sits in a few effective sources, and the assistant estimates move with how contributions are documented and how sources are weighted. Section 5 reports those sensitivities.

The writing flag, by contrast, shows no consistent association with authors, named assistants, or pages. The two indicators also barely overlap: few disclosing papers carry the flag, and few flagged papers disclose. A simple project-choice framework explains why the indicators diverge. Cheaper execution lowers the hours a fixed study needs, but it raises the demand for assistance when researchers widen the study. A saving on writing that is separable from project design leaves inputs unchanged.

Two exploratory splits test that mechanism. Four in five assessable disclosures report coding, data, analysis, or derivation rather than writing alone. Within researcher, those operational disclosures carry the whole assistant difference, 0.679 more named assistants, whereas writing-only statements show none. Operational disclosures also account for the extra length, and most of that length sits in appendices rather than in the main text. The aggregate series tell the same story: in the institutional series the share of papers crediting any assistant stays between 27% and 32% from 2022 to 2026 while AI disclosure rises from zero to about one paper in six. An exploratory

check on multiple-testing corrections, a possible fingerprint of AI in the analysis, proves too rare and noisy to serve as a third indicator.

Voluntary disclosure remains the main threat to this reading. An unattributed team statement dilutes personal stigma across coauthors (Reif, Larrick and Soll, 2025), so the model predicts more authors on disclosing papers without any change in production. But it predicts nothing for assistants or pages. In the data, solo papers disclose more often than coauthored papers in the raw comparison, but that excess vanishes within researcher, and the data reject the account in which each author independently generates a disclosure.

The paper makes two contributions. First, it brings acknowledged junior labour into the economics of generative AI and scientific production. Brynjolfsson, Chandar and Chen (2026) and Hosseini and Lichtinger (2026) document lower junior hiring in exposed jobs, whereas Humlum and Vestergaard (2026) find small average labour-market effects so far. In economics research, operational AI accompanies wider scope and more credited assistants (Acemoglu and Restrepo, 2019; Filimonovic, Rutzer and Wunsch, 2025). Closest in setting, Jiang and Wang (2026) find across economists that AI narrowed the productivity gap between researchers with and without research assistants. Within the same economist's portfolio, disclosing projects credit more assistants. Substitution for an absent assistant and expansion with a present one are compatible, and the two studies observe different margins of the same technology. Steady institutional assistant shares and the operational association are consistent with a junior position that persists on paper, which moves the apprenticeship question in Garicano and Rayo (2025) and Wang (2026) towards the content of the tasks.

Second, the paper separates operational disclosures from prose resemblance in one corpus and treats the disclosure as a collective decision. Kusumegi et al. (2025) and Wang et al. (2026) link detected AI prose to output and to smaller author teams; Hosseinioun, Uzzi and Fosse (2026) find larger teams in AI grant proposals. Estimating both indicators together shows that the writing flag yields intervals that include zero, so modest reductions of the kind Wang et al. report are not ruled out, whereas operational disclosures carry the assistant difference. Observing assistants beside an author difference that is not robust separates delegated execution from coauthorship in theories of knowledge hierarchies and team size (Jones, 2021; Garicano, 2000; Ide and Talamàs, 2025). Modelling self-reported AI use under shared attribution connects the measurement of AI in science to voluntary and group disclosure (Dye, 1985; Levy, 2007; Onuchic and Ray, 2023; Onuchic and Ramos, 2025).

The rest of the paper is structured as follows. Section 2 sets out the project-choice framework that links execution costs and separable writing to project scope and assistant hours. Section 3 describes the 24 series, the assistance and AI measures, the adoption trends, and the multiple-testing check. Section 4 presents the within-researcher specification and the attribution model. Section 5 reports the main estimates, the task-mix and page decompositions, and the sensitivity checks. Section 6 turns to annual researcher portfolios and the training role of research assistance, and Section 7 concludes.

2 Project choice, assistant hours, and separable writing

A research team chooses both the authors on a paper and the breadth of the investigation they carry out. Let $n \geq 1$ denote the number of byline authors and $x \geq 0$ the project's scope: the extent of empirical and analytical work undertaken. In practice, wider scope means assembling extra datasets, running further specifications, checking edge cases, or building fuller appendices. Combining expertise raises a project's value. Execution, verification, coordination among coauthors, and writing consume resources:

$$U(n, x) = B(n)\log(1 + x) - c_o(a_o)x - \frac{jx^2}{2n} - K(n) - C_w(a_w). \tag{1}$$

Here $B(n)$ is the benefit of coauthor expertise and $c_o(a_o)$ is the unit execution cost, the resources spent per unit of routine operational work under AI capability a_o, with $c_o'(a_o) < 0$. Judgement costs $jx^2/(2n)$, with $j > 0$, rise with scope and are shared across n authors. $K(n)$ captures coordination and opportunity costs, and $C_w(a_w)$ is the cost of writing.

The team chooses n and x to maximise $U(n, x)$ and undertakes the paper if the surplus clears its outside option (Appendix C). At a fixed author count, stronger operational AI lowers $c_o(a_o)$ and raises the chosen scope $x^*(n)$, because extra analyses cost less to run. Cheaper execution thus changes what researchers investigate, much as lower production costs widen firms' product lines (Babina et al., 2024) and new tasks offset automation (Acemoglu and Restrepo, 2019). A broader study also raises the marginal value of expertise and spreads the judgement work across coauthors (Garicano, 2000; Ide and Talamàs, 2025). Scope and team size therefore move together, although an integer roster changes only at thresholds and often stays fixed.

Research assistants enter on the delegated execution margin behind the byline. Suppose each unit of scope takes $e(a_o)$ hours of routine assistant work, with $e'(a_o) < 0$, so that delegated assistant hours on the project are $h = e(a_o)x$. Within a smooth choice region for scope,

$$\frac{dh}{da_o} = e'(a_o)x + e(a_o)\frac{dx}{da_o}.$$

The first term is the direct saving on existing tasks: an assistant with a language model cleans data or writes code faster. The second term is the new work created when cheaper execution leads the team to widen the study. Either can dominate.

Hours therefore rise only when the study expands proportionally more than the hours per unit of scope fall. Even the sign of hours does not pin down the number of assistants named on the finished paper. One continuing assistant can absorb the expanded work in more hours, leaving the headcount unchanged. A wider investigation that adds a data source or an estimation step can instead bring a new person into the acknowledgements.

Writing assistance a_w behaves differently when the writing cost $C_w(a_w)$ is separable, meaning that it enters additively and depends on neither scope nor author count. Under separability the writing cost cancels from every comparison between candidate designs, as when an author edits or

translates the prose after the empirical work is complete. Lower writing costs may help a marginal project clear the outside option, but conditional on production they leave the chosen scope and the human inputs unchanged. Separability fails if drafting shapes the argument itself, as Oatley and Djikic (2008) propose, or if writing effort grows with scope and coordination. Polishing the prose of a settled result is separable in this sense; delegating the first formulation of the argument is not (Fourie, 2026). Operational AI thus cuts assistant hours on a fixed study yet can raise the demand for assistance by expanding the study, whereas separable writing help leaves staffing unchanged.

Working papers do not display assistant hours or AI capability directly. AI enters the data only when the authors report using a tool, and human assistance enters only when the acknowledgements credit the work. Authors may disclose a tool when it performed a substantive operation or an outlet asked for a statement, and stay silent when reputational concerns outweigh their reporting norms. Teams that keep fuller records or write longer acknowledgements can also document both AI use and assistant names more often at unchanged staffing. Recorded associations therefore combine project design with reporting choices. For that reason the results condition on acknowledgement documentation, and the empirical approach models the team's decision to disclose.

3 Working papers, acknowledgements, and AI indicators

3.1 Sources

I assemble working papers from 24 series: 23 institutional series and the economics arXiv collection. University departments, central banks, research networks, and policy institutions across North America, Europe, Japan, and Australia supply the institutional outlets. Public catalogues were accessed on 1 November 2026, and cover January 2015 to 31 October 2026, with arXiv entering in September 2017. To prevent selective retrieval, each series had to reach at least 98% public acquisition overall and in each year from 2023 to 2026. Eight series fell short and are excluded: Central Bank of Chile, Pretoria, Chicago BFI, Paris School of Economics, Tokyo CIRJE, Toronto, Yale Cowles, and Stanford SIEPR.

Bank of England and Philadelphia Fed, each missing a single 2026 paper, are retained as the two exceptions. Across the retained series, the source frame contains 56,585 selected paper-family records; Appendix A lists every series. File hashes and corroborated title and author information group versions and cross-postings into practical paper families without guaranteeing complete deduplication. Unavailable PDFs and unresolved text extractions remain missing on the affected measures rather than receiving imputed zeros.

Archived working-paper PDFs can be overwritten by later revisions. Cohort trends and the post-2022 regressions therefore use timing-supported papers, a subset that combines independently version-dated files with documents whose printed dates match their issue cohorts, under the assumption that no unrecorded revision altered the measured text. Papers with conflicting dates or later-revision notices stay outside this subset, and I examine the stricter independently dated subset separately.

3.2 Credited assistance, authors, and pages

Acknowledgements supply two measures of junior labour: whether a paper explicitly credits any research assistance, and the count of named assistants. Named counts record people whose research-assistance role the passage supports, which separates assistants from commenters, advisers, seminar participants, and administrative staff. Collective credits that thank unnamed assistants as a group count as "any" assistance but receive no invented number, so the named count is missing where names are unresolved. AI tools thanked or described as assistants are excluded from the human counts. Recording zero assistance requires readable acknowledgement text rather than an empty or failed extraction. Both outcomes measure credited contributions to a paper, not hours, pay, headcount across papers, or employment.

Author counts come from the printed byline. Pairing the byline with the acknowledgements separates coauthorship from delegated work: a solo author can draw on several assistants, and a large author team can credit none. Page counts record all pages in the retrieved PDF, including references and appendices. Length reflects how much analysis and supplementary material a paper presents, although layout density varies across series and pages do not measure quality.

3.3 Two AI indicators

Two indicators capture distinct ways in which generative AI becomes visible in a working paper. The writing flag applies the released strict *econ-ai-detector* (Goldsmith-Pinkham, 2026) to a consistently selected introduction and body section. The detector is a classifier trained to separate pre-ChatGPT economics articles from language-model text, combining logistic and neural thresholds. Jabarian and Imas (2025) show that detector error rates depend on the threshold and on the text being scored. The strict flag therefore measures resemblance to AI-generated economics prose, not verified use. Short or non-English texts are left missing.

The disclosure indicator records an explicit statement that the authors used generative AI in producing the paper. Candidate passages are retrieved from the full text and classified in context against literal source evidence (Ludwig, Mullainathan and Rambachan, 2026). The coding separates own use in producing the paper from studying AI as a topic, applying conventional tools, leaving attribution ambiguous, or stating non-use. Every positive record, and every candidate surfaced by a corpus-wide audit, received a second review against whole-document passages (Appendix A). No confirmed statement in assessable text forms the reference category, which can contain both undisclosed use and missed statements.

Disclosures also say what the tool did. A disclosure is *writing-only* when every supported task is drafting, editing, translation, or typesetting, and *operational* when any passage reports coding, data preparation, search, analysis, or derivation. About four in five assessable disclosures are operational, and most of the rest are writing-only. Most economists who disclose AI are reporting what it did to their data or their code, not to their prose.

Post-2022 papers observed on both measures show how weakly prose resemblance and explicit disclosure overlap. Only about one in seven disclosing papers carries the writing flag, and about

one in eight flagged papers discloses AI use. The two occur together only modestly more often than they would by chance. Data cleaning or coding leaves no prose trace in an introduction, and light editing need not prompt a disclosure. The regressions therefore enter the writing flag and the disclosure together rather than combining them into one adoption index.

3.4 Adoption since 2022

Figure 1 tracks half-year cohorts of timing-supported papers since 2022. In the institutional series, the share of papers crediting a research assistant holds between 27% and 32% while AI disclosure rises from zero to about one paper in six. Across all sources, disclosed AI use climbs from almost nothing in early 2023 to about one paper in five in the July–October 2026 cohort, which covers only four months because the data end in October; Appendix Table A2 reports the full table.

Rapid disclosure has not coincided with a drop in acknowledged assistance inside the institutional series, or at NBER. The pooled share across all sources does drift down, but outlet composition explains that decline. ArXiv papers rarely credit assistants, and arXiv's share of the sample grows roughly threefold over the period. A possible third indicator, multiple-testing corrections that language models may suggest during an analysis, proves too rare and too noisy to use; Appendix B.8 reports that check.

4 Within-researcher comparison and the disclosure decision

4.1 Comparing papers by the same economist

Economists differ persistently in grant funding, prose style, and the habit of thanking junior staff. A cross-sectional comparison of papers with and without AI indicators would mix those personal traits with project-level differences. I therefore compare papers within the same recurring economist's portfolio, following organisational designs that absorb stable individual practices (Bertrand and Schoar, 2003). *Fixed effects*, separate intercepts for each researcher, each outlet-by-quarter cell, and each broad JEL field, hold constant an author's baseline staffing together with outlet cohort norms and field conventions. What remains is the variation in assistant credits, disclosure, and the writing flag across distinct papers written by the same scholar.

For paper p, linked researchers $i \in \mathcal{A}_p$, working-paper source s, issue quarter q, and broad field f, I estimate

$$Y_p = \alpha_i + \lambda_{sq} + \mu_f + \beta_W W_p + \beta_D D_p + \theta_1 \log L_p + \theta_2 \log K_p + \varepsilon_{pi} \tag{2}$$

on timing-supported papers issued from January 2023 to October 2026. Here Y_p is a paper outcome, W_p is the strict writing flag, and D_p is an explicit disclosure of generative AI use. Both indicators enter together rather than as a combined adoption index. The terms $\log L_p$ and $\log K_p$ record selected prose words and scored text windows, which hold constant the detector's opportunity to register a signal. Each paper enters once per included linked researcher with row weight $1/|\mathcal{A}_p|$, so

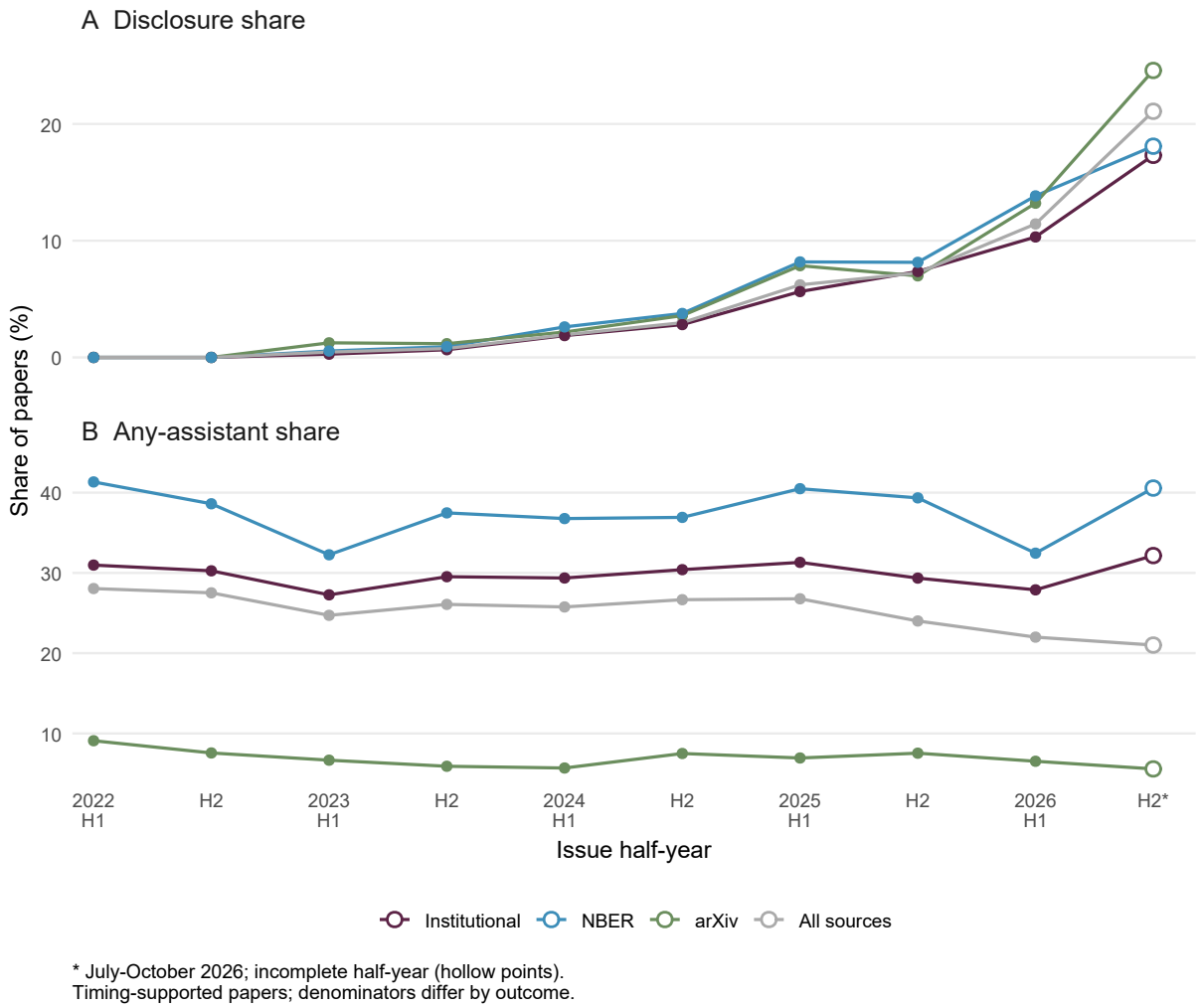


Figure 1. AI disclosure and credited assistance

Half-year cohorts of timing-supported papers. Panel A shows the disclosure share; Panel B shows the share crediting any assistant. The last point covers only July–October 2026, because the data end in October. Denominators differ by outcome; Appendix A reports shares and counts.

every paper has a total weight of one.

Requiring usable text, linked identities, and at least two post-2022 papers per researcher narrows the timing-supported records considerably. After the iterative pruning of singletons, the assistant models retain 7,243 papers by 4,604 recurring researchers, 468 of them disclosing AI use; the byline and length models retain 8,502 papers by 5,285 researchers, 544 of them disclosing. Appendix A traces the sample from the 56,585-record frame, and Appendix B reports pooled models without researcher fixed effects on identical samples.

Standard errors cluster by researcher and by paper, which lets residuals correlate across a scholar's portfolio and across the coauthor rows of one manuscript. But the residual disclosure variation sits in the equivalent of about 3.3 equally informative sources, and shared coauthors link

papers across series without sharing a researcher intercept. Source-quarter intercepts do not absorb source-wide shifts in disclosure norms or cross-outlet coauthor covariance; Appendix B.1 evaluates both.

Across the registered family of twenty tests, two indicators across ten outcomes, I lead with Romano–Wolf step-down p-values from a researcher-level wild cluster bootstrap and report the registered Holm adjustment beside them. The Romano–Wolf procedure asks how often, across many re-drawn datasets in which no indicator matters, the largest test statistic would be as large as the one observed. It allows for the correlation among the twenty tests that Holm ignores.

The coefficients β_W and β_D measure conditional associations across projects within an economist's recent output. Neither estimates a before-and-after change on a fixed investigation, and a scholar's first flagged or disclosing paper is never treated as an adoption date. Defining treatment timing from an observed output feature would induce mechanical pre-trends, because prolific authors cross any threshold earlier (Renault, Bergeaud and Bosquet, 2026). Within a portfolio, an economist can deploy AI on an ambitious empirical study, omit it on a short note, or report it selectively. Separating scope expansion from selective reporting requires a model of why a team chooses to disclose.

4.2 Disclosure as a decision

A detector score is computed from the text. A disclosure is a public statement chosen by the authors after the work is done. Reif, Larrick, and Soll (2025) document that observers penalise workers who use AI assistance, which gives scholars a reputational reason to hesitate before reporting a tool. On a solo paper, an acknowledgement of AI use identifies the one author as the operator. On a coauthored paper, an unattributed statement reveals that someone on the team used AI without naming who ran the model. Coauthorship can therefore dilute personal attribution even when the technology changes no team, much as delegation shifts responsibility in other settings (Bartling and Fischbacher, 2012; Chassang and Zehnder, 2024).

Consider completed projects with a predetermined team of $n \geq 1$ authors that used AI with a fixed probability across team sizes. Conditional on use, one author operated the tool with prior probability $1/n$, and silence leaves readers with posterior probability r_n that a silent paper used AI. Every author bears an undivided personal cost $b \geq 0$ from withholding a truthful statement, which can stand for honesty norms or an outlet's expectations, and a reputational loss of $s > 0$ times the reader's posterior probability that this author operated the tool. An unattributed disclosure assigns operator probability $1/n$ and silence assigns r_n/n, so each author's net payoff from disclosing rather than staying silent is

$$\Delta u_n = b - \frac{s}{n}(1 - r_n). \tag{3}$$

Appendix C.3 derives the unique rational-expectations cutoff equilibrium and shows that the disclosure threshold on b falls strictly as n rises. The personal cost of silence stays at b for every collaborator, whereas the incremental stigma $(s/n)(1 - r_n)$ shrinks with team size.

Under homogeneous preferences, the model predicts that disclosing papers average more authors than silent papers even when AI leaves every roster unchanged. Across assessable post-2022 papers, disclosing papers average only marginally more authors than silent ones, and both medians equal three. For the main estimates the important point is what the model does not predict. It supplies no direct prediction for research assistants or page length. Attribution selection could generate assistant or page differences only if larger author teams already used more assistants and wrote longer papers while the rest of the staffing was unrelated to reporting. Heterogeneous attitudes and a unanimity rule can even reverse the author gradient, because a larger team contains more possible objectors.

Figure 2 compares raw disclosure rates across team sizes with within-researcher differences relative to solo papers. In the raw data, solo papers disclose AI use more often than papers with any larger team. That solo excess contradicts the homogeneous model's pooled prediction. But it disappears once researcher fixed effects enter. Within the same economist's portfolio, no team size differs significantly from solo papers in its disclosure probability, and a doubling of the author count is associated with no detectable change.

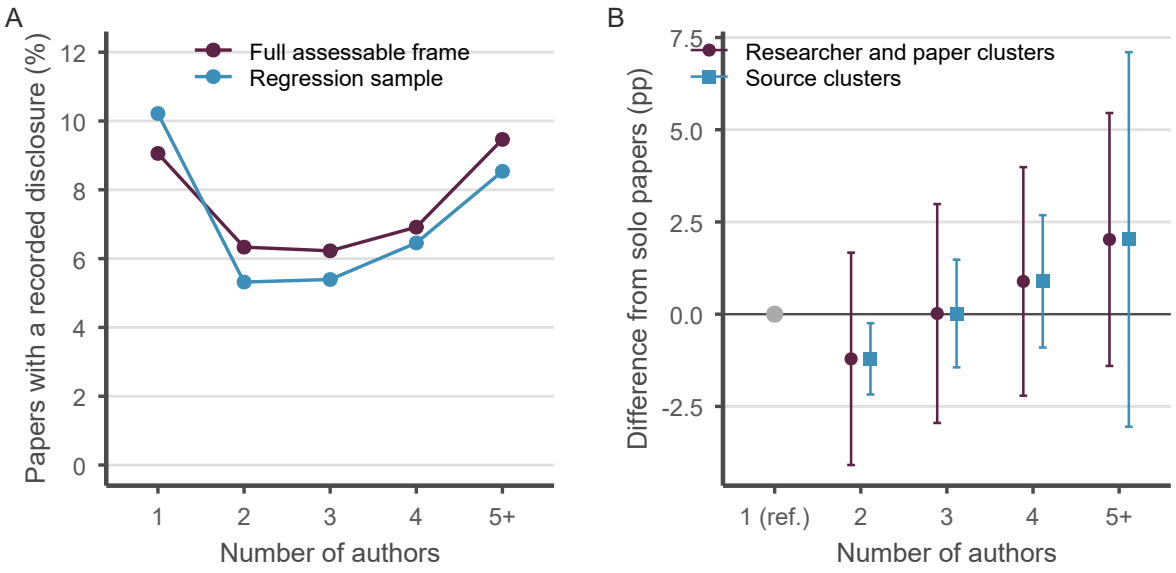


Figure 2. Recorded disclosure by number of authors

Panel A: share of timing-supported post-2022 papers with a recorded own-use disclosure, by author count, in the full assessable frame (16,300 papers, 1,167 disclosures) and in the author-count regression sample (8,502 papers, 544 disclosures). Panel B: within-researcher differences in disclosure probability relative to solo papers, in percentage points, from a linear probability model with researcher, source-quarter, and field effects, the writing flag, and prose controls, on the 8,502-paper sample; 95% intervals clustered by researcher and paper (circles) and by source (squares). Disclosure information is concentrated in about four effective sources. No recorded statement does not establish non-use.

A mechanical benchmark offers a separate check on how disclosures aggregate across coauthors. If each author independently triggered a disclosure with the same probability within a source-year

cell, a complementary log-log regression, a link function under which independent union probabilities are linear in log team size, would give a slope of one on $\log n$. The estimated slope is far below one. The data reject identical independent author-level reporting without ruling out heterogeneous task roles. Before estimation, I also froze eight competing mechanism tests, reported in Appendix B.6: coauthor field overlap and citation standing, coauthors' pre-2022 documentation habits, outside-team propensities, detection-recall bounds, submission-preparation signals, and protective wording (Appendix B.5). Coauthored disclosures use protective wording somewhat more often, but no test survives the Holm adjustment. The estimable tests rest on very few effective sources and separate neither attribution nor documentation habits from actual AI use.

5 AI indicators, research assistants, and project scope

5.1 Main estimates

Table 1 reports within-researcher estimates from Equation (2) for the two AI indicators across four primary outcomes: any credited research assistant, the count of named assistants, the count of byline authors, and total PDF pages. Both indicators enter together with the prose controls and the researcher, source-quarter, and field fixed effects. Papers without a disclosure in the estimation sample average 0.83 named assistants and 52.4 pages. Reading across the two columns asks whether an explicit AI statement and AI-like prose move with human inputs in the same way.

Table 1. Within-researcher associations between AI indicators and paper inputs

Outcome	Writing flag	Disclosed	Papers	Disclosed papers	Researchers
Any assistant (pp)	-3.206	6.599	7,243	468	4,604
	(2.251)	(2.606)			
Named assistants	-0.131	0.548	7,242	468	4,603
	(0.111)	(0.168)			
Authors	0.011	0.240	8,502	544	5,285
	(0.061)	(0.093)			
Pages	-0.693	5.346	8,502	544	5,285
	(1.100)	(1.274)			

Estimates are from Equation 2 with the writing flag and disclosure entered together, prose-length and window controls, researcher, source-quarter, and field fixed effects, and paper-total weights. Standard errors clustered by researcher and paper in parentheses; any-assistant estimates are in percentage points. Across the twenty-test family (two indicators by ten outcomes), the disclosure p-values are: any assistant raw p = .011, Romano–Wolf p = .157, Holm p = .170; named assistants raw p = .001, Romano–Wolf p = .020, Holm p = .019; authors raw p = .010, Romano–Wolf p = .147, Holm p = .163; pages raw p < .001, Romano–Wolf p = .0005, Holm p = .0006. All ten writing-flag tests have Romano–Wolf p above .64. Neither adjustment accounts for the 3.3 effective sources.

Within an economist's portfolio, papers that disclose AI use credit more named research assistants and run longer. Disclosing papers credit 0.548 additional named assistants, two-thirds of the mean on papers without a disclosure, and the association survives the twenty-test adjustment

under both Romano–Wolf and Holm. They also run 5.35 pages longer, about a tenth more than the baseline, an even more precisely estimated difference. Disclosing papers are also more likely to credit any assistant and carry slightly more authors. Both differences are individually significant, but neither survives the family-wise adjustment, and neither carries weight in what follows.

Resemblance to AI-generated writing shows a different pattern in the first column. Flagged papers differ little from unflagged papers by the same economist on any of the four outcomes, and every interval includes zero. In pooled regressions without researcher intercepts (Appendix Table B2), flagged papers credit noticeably less assistance; comparing projects within the same economist shrinks that gap toward zero. Read through Equation (1), the two columns fit a technology that expands scope when applied to research operations and leaves staffing untouched when confined to separable prose editing.

5.2 Which disclosures carry the association

If scope expansion explains the positive assistant estimate in Table 1, disclosures that report research operations should carry the association and statements confined to prose editing should look like the writing flag. I test that implication in an exploratory split outside the registered twenty-test family and report raw p-values. Using the whole-document task rule described in the data section, disclosures divide into operational statements, writing-only statements, and unlabelled statements, the last controlled separately. Operational disclosures outnumber writing-only ones by about five to one in every sample, so the writing-only estimates are less precise. Table 2 reports the within-researcher coefficients and a test of equality between the operational and writing-only coefficients for the four main outcomes.

Table 2. Exploratory within-researcher estimates by the disclosed AI task mix

Outcome	Operational	Writing-only	Equality p
Any assistant (pp)	6.9 (2.9)	1.7 (6.0)	.43
	.020	.78	
Named assistants	0.679 (0.199)	-0.098 (0.209)	.0034
	.0007	.64	
Authors	0.300 (0.103)	0.056 (0.235)	.32
	.0035	.81	
Pages	6.78 (1.44)	0.83 (2.97)	.071
	$< .0001$	.78	

Entries are the coefficient, the researcher/paper-clustered standard error in parentheses, and the raw p-value from replacing the single disclosure indicator in Equation 2 with operational-or-mixed, writing-only, and unlabelled indicators; the final column reports the raw p-value of a test that the operational and writing-only coefficients are equal. Samples match Table 1, with 361 operational, 68 writing-only, and 39 unlabelled disclosures for assistance and 425, 78, and 41 for authors and pages. All tests lie outside the registered twenty-test family.

Operational disclosures account for the whole named-assistant association. Within a researcher's portfolio, papers that report coding, data, analytical, or mathematical AI use credit 0.679 more

named assistants, whereas papers that report only drafting, editing, translation, or typesetting credit no more assistants than the same researcher's silent papers. The two coefficients differ clearly. The contrast matters for the reporting explanation. Writing-only authors also chose to place an AI statement in their paper, so a general willingness to document every input would raise assistant counts in both columns, not only in the operational one.

Document length follows the same division. Operational disclosures run several pages longer within researcher, whereas writing-only disclosures are barely longer than silent papers, although the test of equality is only marginally significant. The raw means tell the same story: operational papers credit the most assistants and run longest, and writing-only papers sit slightly below papers without a disclosure on both. Pooled regressions without researcher intercepts also separate the two types on named assistants and pages, but not on authors or incidence. The writing-only cells are modest, and only a handful of economists have papers in both task categories, so the comparison rests mainly on contrasting each disclosure group with the silent papers of the same scholar.

5.3 Where the extra pages are

Five extra pages on a disclosing paper could be padded main text or additional supporting analyses, robustness tables, and data documentation in an appendix. To locate the added length I run a second exploratory analysis outside the registered family, with raw p-values, on almost all of the pages-sample manuscripts, those with verified page text. The boundary rule takes the first appendix page from usable PDF bookmarks where they exist and otherwise from a standalone appendix heading or a lettered section that follows the last numbered section or the bibliography. On a held-out validation sample, the rule places the appendix boundary correctly in about nine papers in ten; Appendix B.8 reports the error rates. Table 3 splits total length into main-text pages, appendix pages, body pages before both references and appendix, and an indicator for having any appendix.

About seven-tenths of the extra length on disclosing papers sits in the appendix and three-tenths in the main text. Disclosing papers are not much more likely to have an appendix at all; their appendices are longer. Operational disclosures lean even more on the back matter, whereas writing-only disclosures show no appendix gain. The appendix share of the extra length holds when NBER or arXiv is dropped and among papers whose boundaries are confirmed by bookmarks. The writing flag, by contrast, points toward shorter appendices and slightly longer main text, a pattern that is suggestive at most.

5.4 Documentation and source composition

Figure 3 places the main disclosure coefficients beside alternative functional forms, controls for acknowledgement space, and source-weighting schemes. Conditioning on the acknowledgement word count attenuates both assistant estimates modestly on identical samples. Thanking assistants and stating AI use both add words to an acknowledgement, so conditioning on its length absorbs part of the outcome as well as general verbosity. It is a descriptive check, not a bias correction.

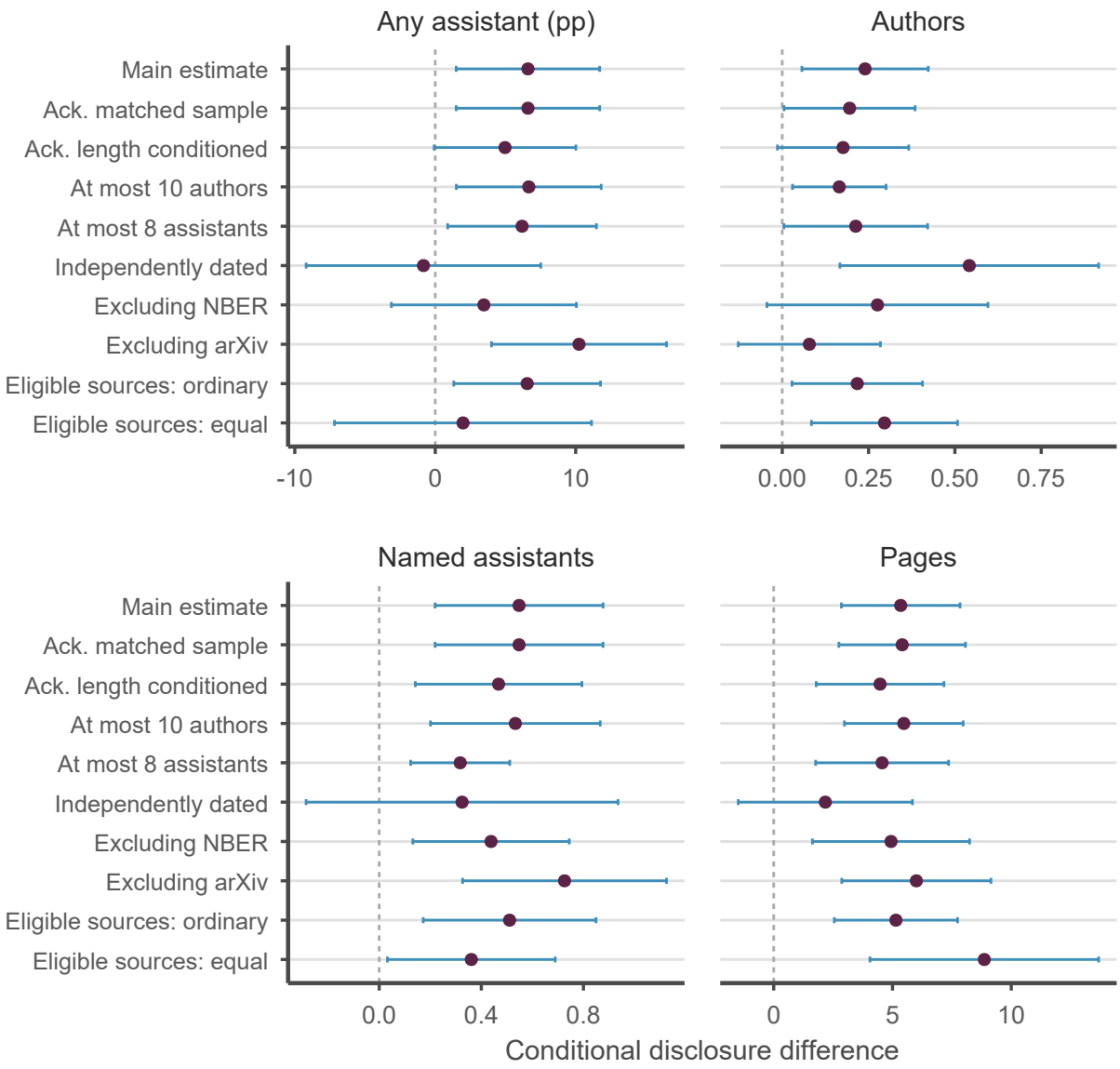


Figure 3. Disclosure, documentation, and source composition

Disclosure coefficients with pointwise 95% researcher/paper-clustered intervals, conditional on that dependence structure. Acknowledgement-length conditioning is compared with its same-sample baseline. Equal source weights are compared with ordinary weights on the same eligible-source subset. Both AI indicators and the fixed effects in equation 2 remain included. Figure 4 uses the same specification rows for three writing indicators and two team outcomes. Source concentration and overlapping authors limit interpretation of the intervals; see Appendix B.1.

Table 3. Exploratory within-researcher decomposition of paper length into main text and appendix

Outcome	Disclosed	Operational	Writing-only
Total pages	5.29 (1.28)	6.70 (1.45)	0.81 (2.97)
	$< .001$	$< .0001$	.79
Main-text pages	1.63 (0.70)	1.70 (0.76)	3.02 (1.83)
	.020	.026	.098
Appendix pages	3.66 (1.10)	5.00 (1.28)	-2.22 (2.27)
	.0009	.0001	.33
Body pages (before references and appendix)	1.16 (0.61)	1.07 (0.67)	2.63 (1.65)
	.059	.11	.11
Any appendix (pp)	3.4 (2.2)	2.3 (2.1)	1.9 (4.5)
	.11	.26	.67

Estimates use the 8,493-paper verified-text subset with 542 disclosures (the full-sample total-page estimate in Table 1 is 5.35), the Equation 2 controls and fixed effects, researcher/paper-clustered standard errors in parentheses, and raw p-values. Column 1 uses the single disclosure indicator; columns 2 and 3 come from the task-mix specification. All estimates lie outside the registered twenty-test family.

Large or skewed assistant rosters contribute part, though not all, of the headcount magnitude. The log of named assistants and a count capped at eight names both remain strongly significant across the twenty-test family. Dropping papers that credit more than eight assistants shrinks the linear estimate by about two-fifths (Appendix B.2), so ordinary teams and large lab-style rosters both contribute. A separate count of documented empirical and theoretical methods fails the adjustment (Appendix B.4).

Outlet composition shapes binary assistant incidence far more than named headcounts or page length. Omitting NBER makes the incidence coefficient smaller and imprecise, whereas omitting arXiv, where acknowledgements rarely list junior staff, makes it larger. Weighting outlets equally rather than by paper reduces incidence to an imprecise estimate near zero and trims the named-assistant estimate. Dropping any single source leaves the page difference essentially unchanged, while the author difference shrinks when arXiv is dropped. Restricting links to work-corroborated researcher identities keeps all four disclosure coefficients positive with a smaller author estimate. The small independently dated subset gives larger author and smaller page differences, and no incidence difference.

5.5 Writing resemblance

Figure 4 plots within-researcher estimates with pointwise 95% intervals for author and named-assistant counts across ten sensitivity specifications and three text flags: the strict *econ-ai-detector* indicator, a high colon rate, and a high dash rate. Every specification pairs one text indicator with the disclosure and the Equation (2) controls on identical samples. Across the ten specifications, neither the detector flag nor the colon rate shows a consistent association with author or named-assistant counts: every detector and colon interval includes zero. The dash rate gives a negative author association in two specifications, under the ten-author cap and when NBER is omitted; all

other dash intervals span zero.

The intervals in the main detector specification are nevertheless wide enough to permit moderate reductions in staffing, especially in named assistants. The detector estimates therefore do not rule out team contractions of the size Wang et al. (2026) report in other disciplines with different prose classifiers. Measurement error in automated detection (Jabarian and Imas, 2025) and unrecorded AI use among silent papers can both attenuate the prose-flag comparisons. But the contrast with the operational disclosures matches the separable-writing case of Equation (1). The prose a detector reads is a measure of how the paper was written, not of the argument or evidence beneath it (Fourie, 2026).

Appendix A reports supplementary reference-list and university-affiliation comparisons. Disclosing papers cite more recent references, but the gap more than halves once papers about AI itself are excluded, and they cite no more references in total. Conditioning on university-tier shares leaves the any-assistant and page estimates essentially unchanged.

6 Apprenticeship, portfolios, and the division of research labour

Within a recurring economist's portfolio, projects that disclose operational AI credit more named research assistants and run longer than silent papers by the same scholar. The institutional series show a matching aggregate pattern: the share crediting any assistant holds steady in every half-year from 2022 to 2026 as disclosure rises. On the finished manuscript, the junior rung of academic economics has not disappeared with the arrival of language models. But paper-level regressions speak only to acknowledged contributions on completed working papers by recurring scholars. They do not observe department-wide hiring, junior wages, contract durations, or projects abandoned before circulation, and they exclude economists who post too rarely to enter the sample.

The distinction between comparing across economists and comparing within one economist's portfolio reconciles these findings with recent evidence on unequal gains from AI. Jiang and Wang (2026) find that generative AI raised the top-journal publication share of economists who work without research assistants. Substitution across persons and expansion within a portfolio are distinct margins of the same technology. For a scholar working without human support, a language model performs the routine coding and data cleaning that once required grant money to hire an assistant. Within an assisted economist's portfolio, cheaper execution lowers the operational cost in Equation (1) and makes wider studies, with fuller appendices and more credited names, worth undertaking. In Swedish occupational data from 1880 to 2019, Skoglund, Molinder and Enflo (2026) find that most new work arose from Smithian specialization as scale deepened the division of labour, whereas occupations tied directly to a new technology stayed small and grew older. Research assistance is a Smithian role: a wider study generates specialized junior tasks, not a new occupation.

Even inside an assisted research group, more credited names per paper need not mean more junior jobs or more paid hours. An acknowledgement records a person's contribution to a paper, not an employment contract or a time sheet. One full-time pre-doctoral fellow can spend a year on

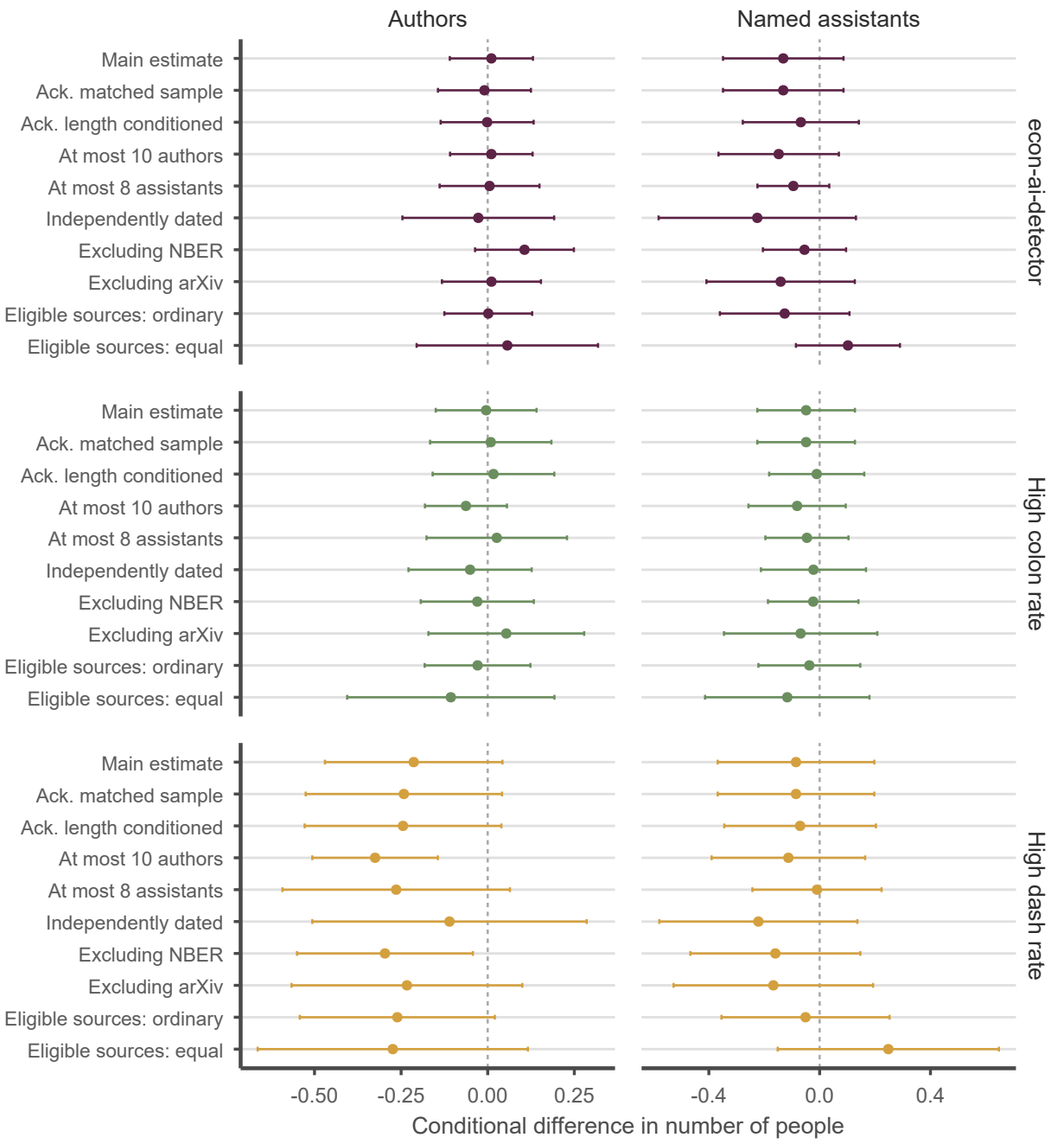


Figure 4. Writing indicators, authors, and named assistants

Coefficients with pointwise 95% researcher/paper-clustered intervals. Each writing indicator enters separately with disclosure and the controls and fixed effects in equation 2. Corresponding rows use identical samples and weights across indicators. High colon and dash rates exceed the historical pooled mean plus two standard deviations (8.905 and 9.267 per 1,000 words). Equal-source comparisons retain sources with at least five disclosed and twenty no-statement papers. Scales match within each outcome. These are writing proxies, not verified AI use; specifications overlap and uncertainty remains conditional on the stated dependence structure.

a single paper and next year contribute shorter tasks to three papers, raising credits per paper at unchanged headcount. A wider investigation can also add a part-time student for one specialized data check while total assistant hours in the group fall. Paper-level estimates therefore leave open how an economist's total demand for assistants evolves across an annual portfolio.

Appendix B.7 aggregates manuscripts into complete annual researcher portfolios across the series present throughout the window, without conditioning on disclosure. Mean credited assistants per paper dipped in 2023 and then rose above its 2022 level by 2025. Whether that increase reflects new hiring or wider reuse across papers turns on how many distinct people stand behind a scholar's annual credits. Acknowledgements print names without affiliations or dates, so matching strings across two papers can mean one reused assistant or two people who share a name. Identity-free bounds, set between a scholar's largest single-paper roster in that year and the total annual credits, place distinct assistants per portfolio only modestly above one in 2025. Coauthors' portfolios also overlap on joint manuscripts, so researcher-level counts cannot be summed into a discipline-wide workforce.

The margin these credits cannot see, the flow of new entrants, is where a slowly diffusing technology has adjusted before. In English bootmaking after the sewing machine, Vipond (2026) finds that about 153,000 artisanal jobs disappeared and about 140,000 specialized jobs emerged, yet incumbent male artisans were neither displaced nor moved into the new work. Young men stopped entering the old trade, and the new jobs went to young workers in other places. Schneider and Vipond (2023) argue that past transitions often took decades and that access to the new work was stratified by gender, age, education, and location. Four years of paper-level credits therefore cannot rule out a collapse of entry.

Task allocation inside those portfolios matters beyond short-run headcounts because research assistance is a central apprenticeship of economics. Novices clean raw datasets and write estimation code in exchange for supervision and for letters that carry weight in doctoral admissions. Once AI absorbs enough routine work that doing the task oneself costs a senior less than instructing a novice, the surplus that finances mentorship disappears and the apprenticeship breaks down (Garicano and Rayo, 2025). Emanuel, Harrington, and Pallais (2026) find that proximity to experienced coworkers raises the feedback junior workers receive.

Steady institutional assistant shares and the operational disclosure association are consistent with a junior position that persists on paper. Credits cannot reveal where that position sits relative to the viability threshold. Identical thank-you lines can cover different bundles of junior work as job boundaries shift. On one disclosing project an assistant may link archival sources and run supplementary estimations that cheaper computation made worth pursuing. On another the assistant may check model-generated scripts and audit appendix numbers produced from a senior author's prompts (Agrawal, McHale and Oettl, 2026). Building an empirical pipeline from raw files teaches a novice to spot data anomalies and identification threats; reviewing fluent machine output without writing the code may teach verification without that foundation.

7 Conclusion

Working papers that disclose generative AI use credit more named research assistants and run longer than other papers by the same economist. Disclosures of coding, data preparation, analysis, or derivation carry both differences, and most of the added length sits in appendices. Disclosures of writing help alone and AI-like prose show no consistent association with authors, named assistants, or pages. Among the projects that recurring economists circulate as working papers, AI used for research operations accompanies wider, assistant-supported investigations rather than assistantless ones.

These findings qualify the evidence that AI is displacing entry-level knowledge workers. Payroll and job-posting studies find slower entry-level hiring in occupations where AI automates routine tasks. Academic economics records junior labour next to finished output, and there the institutional series pair rising AI disclosure with a steady share of papers crediting assistants and with larger rosters on projects that use AI for research operations. Project-level records reveal how teams widen their projects when AI lowers the cost of execution, a reorganisation within the workflow that employment counts by occupation miss. The weak overlap between the two AI measures also carries a lesson for measuring AI at work: a voluntary disclosure and prose that resembles AI output are different objects.

The findings carry two implications for workplace governance and training. First, universities, journals, and knowledge-intensive firms increasingly rely on self-reported AI use to track adoption. Here, voluntary disclosures record operational workflows and barely overlap with AI-like prose, so neither self-reports nor text detectors alone offer a complete audit of AI use at work. Second, steady assistant credits alongside wider projects suggest that, for now, the pressing question for entry-level knowledge work concerns the content of junior tasks more than headcounts. Where AI absorbs first-pass coding and data work, staffing juniors on larger projects need not build foundational skills, and supervisors, employers, and the programmes that fund assistantships may need to pair the checking of machine output with tasks in which novices build an analysis themselves.

The evidence describes credited contributions to completed papers in 24 working-paper series, not paid hours, wages, distinct headcounts, or department-wide employment, and the disclosure variation is concentrated in a few series. Comparing papers by the same economist and contrasting operational with writing-only statements narrow the simplest documentation explanations, but unobserved project ambition and selective silence among papers without a disclosure remain part of every comparison.

Before a working paper reaches a series, a language model can now clean the raw files, write the estimation code, search the literature, and assemble the appendix tables in minutes. When AI takes that first pass, does the research assistant thanked behind the byline build broader evidence or check machine output? And, more importantly, does checking still teach a novice how to do economics?

References

Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives*, 33(2): 3–30.

Agrawal, Ajay K., John McHale, and Alexander Oettl. 2026. "AI in Science." National Bureau of Economic Research 34953.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2024. "Artificial Intelligence, Firm Growth, and Product Innovation." *Journal of Financial Economics*, 151: 103745.

Bartling, Björn, and Urs Fischbacher. 2012. "Shifting the Blame: On Delegation and Responsibility." *Review of Economic Studies*, 79(1): 67–87.

Bertrand, Marianne, and Antoinette Schoar. 2003. "Managing with Style: The Effect of Managers on Firm Policies." *Quarterly Journal of Economics*, 118(4): 1169–1208.

Brynjolfsson, Erik, Bharat Chandar, and Ruyu Chen. 2026. "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence." Stanford Digital Economy Lab. Working paper, August 2025, revised August 2026.

Chassang, Sylvain, and Christian Zehnder. 2024. "Secure Survey Design in Organizations: Theory and Experiments." *American Economic Journal: Microeconomics*, 16(4): 371–405.

Dye, Ronald A. 1985. "Disclosure of Nonproprietary Information." *Journal of Accounting Research*, 23(1): 123–145.

Emanuel, Natalia, Emma Harrington, and Amanda Pallais. 2026. "The Power of Proximity to Coworkers." *Quarterly Journal of Economics*, 141(3): 1825–1870.

Filimonovic, Dragan, Christian Rutzer, and Conny Wunsch. 2025. "Can GenAI Improve Academic Performance? Evidence from the Social and Behavioral Sciences." arXiv 2510.02408.

Fourie, Johan. 2026. "When Everyone Writes Well: The Economics of Writing with AI." Department of Economics, Stellenbosch University Working Paper.

Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy*, 108(5): 874–904.

Garicano, Luis, and Luis Rayo. 2025. "Training in the Age of AI: A Theory of Career Viability." Centre for Economic Policy Research Discussion Paper 20634. Revised March 2026.

Goldsmith-Pinkham, Paul. 2026. "econ-ai-detector." Software repository; pinned release described in Appendix A.

Hosseini, Seyed M., and Guy Lichtinger. 2026. "Generative AI as Seniority-Biased Technological Change: Evidence from U.S. Résumé and Job Posting Data." SSRN 5425555. Posted September 2025, revised May 2026.

Hosseinioun, Moh, Brian Uzzi, and Henrik Barslund Fosse. 2026. "Artificial Intelligence in Science: Returns, Reallocation, and Reorganization." arXiv 2603.27956.

Humlum, Anders, and Emilie Vestergaard. 2026. "Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI." National Bureau of Economic Research 33777. Revised March 2026.

Ide, Enrique, and Eduard Talamàs. 2025. "Artificial Intelligence in the Knowledge Economy." *Journal of Political Economy*, 133(12): 3762–3800.

Jabarian, Brian, and Alex Imas. 2025. "Artificial Writing and Automated Detection." SSRN 5407424.

Jiang, Tianyu, and Yuqing Wang. 2026. "Equalizing or Polarizing? Generative AI and Research Productivity in Economics." Job market paper, University of Michigan.

Jones, Benjamin F. 2021. "The Rise of Research Teams: Benefits and Costs in Economics." *Journal of Economic Perspectives*, 35(2): 191–216.

Korinek, Anton. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature*, 61(4): 1281–1317.

Korinek, Anton. 2025. "AI Agents for Economic Research." National Bureau of Economic Research 34202.

Kusumegi, Keigo, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. 2025. "Scientific Production in the Era of Large Language Models." *Science*, 390(6779): 1240–1243.

Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review*, 97(1): 150–168.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2026. "Large Language Models: An Applied Econometric Framework." *Annual Review of Economics*, 18: 283–316.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on Productivity Effects of Generative Artificial Intelligence." *Science*, 381(6654): 187–192.

Oatley, Keith, and Maja Djikic. 2008. "Writing as Thinking." *Review of General Psychology*, 12(1): 9–27.

Onuchic, Paula, and Debraj Ray. 2023. "Signaling and Discrimination in Collaborative Projects." *American Economic Review*, 113(1): 210–252.

Onuchic, Paula, and João Ramos. 2025. "Disclosure by Groups." Centre for Economic Policy Research Discussion Paper 20057.

Reif, Jessica A., Richard P. Larrick, and Jack B. Soll. 2025. "Evidence of a Social Evaluation Penalty for Using AI." *Proceedings of the National Academy of Sciences*, 122(19): e2426766122.

Renault, Thomas, Antonin Bergeaud, and Clément Bosquet. 2026. "Scientific Production in the Era of Large Language Models: Outcome-Triggered Treatment Timing and Spurious Event-Study Dynamics." *Proceedings of the National Academy of Sciences*, 123(33): e2618638123.

Schneider, Benjamin, and Hillary Vipond. 2023. "The Past and Future of Work: How History Can Inform the Age of Automation." CESifo Working Paper 10766. November 2023.

Skoglund, William, Jakob Molinder, and Kerstin Enflo. 2026. "What the History of (New) Jobs Can Tell Us about the Impact of Technological Change on Employment." Lund Papers in Economic History 2026:273, Lund University, September 2026.

Vipond, Hillary. 2026. "Technological Unemployment in Victorian Britain: Young Workers and the Collapse of Entry." University of Chicago Stone Center for Research on Wealth Inequality and Mobility Working Paper 26-13. August 2026.

Wang, Haoyang, Mingze Zhang, Yi Bu, Star Xing Zhao, and Meijun Liu. 2026. "Smaller, Younger, and More Impactful: How AI-Assisted Writing Transforms Research Teams." arXiv 2605.27404.

Wang, Shaohui. 2026. "Hope, Signals, and Silicon: A Game-Theoretic Model of the Pre-doctoral Academic Labor Market in the Age of AI." *Research Policy*, 55(8): 105568.

A Data, measures, and samples

A.1 Source coverage and document timing

Institutional screening considered established RePEc-listed outlets, geographical coverage, and annual catalogue volume. Annual catalogue volume, including a ten-item benchmark, was a screening preference rather than a uniform eligibility threshold. Selected-family and regression counts can be smaller. Penn PIER is excluded because its retained post-2022 coverage consists of one paper. The accessible Swiss Finance Institute copies do not provide coverage on both sides of 2022. The numbered Bank of England Staff Working Papers and Boston Fed Research Department Working Papers are reconciled within the sample window. Publisher withdrawals and unavailable source files retain explicit missing-source dispositions. Catalogue enumeration does not establish that every paper has retrievable, date-supported text. The Central Bank of Chile, Pretoria, Chicago BFI, Paris School of Economics, Tokyo CIRJE, Toronto, Yale Cowles, and Stanford SIEPR series could not be completely acquired from public sources and are excluded to avoid selective coverage. The coverage rule requires at least 98% acquisition overall and in each year from 2023 to 2026. Bank of England and Philadelphia Fed are retained exceptions: their 2026 coverage is 47/48 and 46/47 papers, respectively, each one paper short of complete coverage. The retained series and family counts appear in Table A1. The sources are a selected set of outlets, not a representative sample of economists.

Table A1 separates retained records, supported timing, and observed measures. Acquisition and deduplication mean that retained counts can differ from catalogue counts. In particular, an absent retained family in a source-year is not evidence that the outlet issued no papers. Unavailable PDFs and unresolved extraction remain missing in the affected measure. Exact hashes and corroborated title/author information identify candidate versions and cross-posts; date and source priorities choose representatives without replacing a record because of a missing outcome. Residual cross-posts can remain, although no fitted outcome contains two records with an identical PDF.

The frame comprises 11,160 independently version-dated documents, 31,303 additional documents included under the printed-date assumption, and 14,122 with unresolved or later-revision timing. The first two groups constitute the 42,463 timing-supported records. An independently dated copy supplies evidence about when the measured text existed. A printed date consistent with its issue cohort instead requires no unrecorded later revision affecting the measures; a file hash alone cannot establish this. Explicit later revisions, conflicting dates, and undated copies do not automatically supply historical outcomes. Repository citation sheets do not substitute for dates on manuscript front matter. A hypothetical publisher notice that a paper may be revised is not evidence of an actual revision. A bare manuscript-version date remains a printed-date observation; it does not establish an independently verified first edition. Actual revision notices and later front-matter dates retain the same timing restrictions. Figure A1 ends with July–October 2026, an incomplete half-year.

Table A1. Coverage by retained source

Series	Records	Dated	Scored	Assistance	Disclosed
arXiv economics	17,204	7,332	12,824	11,288	1,010
Bank of Canada	641	599	617	628	4
BIS	823	699	741	771	19
Bank of England	656	541	594	606	3
Warwick CAGE	588	514	491	553	15
Cambridge	929	772	786	770	14
US Census CES	654	355	485	577	3
LSE CEP	709	611	648	702	17
CESifo	7,284	6,377	6,586	6,066	172
Osaka ISER	421	366	367	385	11
ANU CAMA	840	798	763	667	8
Boston Fed	212	139	181	195	2
San Francisco Fed	282	172	232	252	4
Federal Reserve Board	1,072	807	908	943	14
Chicago Fed	249	153	206	214	9
Institute for Fiscal Studies	358	249	326	341	10
NBER	14,836	13,876	11,723	13,919	345
New York Fed	477	295	410	432	5
Philadelphia Fed	425	378	355	376	4
RIETI	1,256	1,144	1,058	1,246	8
Tinbergen	968	872	894	780	10
Toulouse	974	846	817	840	7
Pompeu Fabra	414	279	340	371	4
World Bank	4,313	4,289	3,479	3,750	31
Total	56,585	42,463	45,831	46,672	1,729

Records are selected paper-family representatives. Dated denotes timing-supported documents. Scored and assistance count assessable writing and assistant-incidence observations in the full source frame; they need not be timing-supported. Disclosed counts positive statements after the second review described in Section A.4, not documents searched. These columns therefore have different denominators and are not successive stages of a single sample restriction.

Table A2. Half-year disclosure and assistant credits

Group	Half-year	Disclosed (%)	N_D	Assistant (%)	N_A
All sources	2022 H1	0.00	1,734	28.05	1,565
arXiv	2022 H1	0.00	314	9.09	209
Institutional	2022 H1	0.00	1,420	30.97	1,356
NBER	2022 H1	0.00	517	41.33	525
All sources	2022 H2	0.00	1,712	27.52	1,530
arXiv	2022 H2	0.00	281	7.57	185

Group	Half-year	Disclosed (%)	N_D	Assistant (%)	N_A
Institutional	2022 H2	0.00	1,431	30.26	1,345
NBER	2022 H2	0.00	551	38.61	562
All sources	2023 H1	0.45	1,766	24.71	1,574
arXiv	2023 H1	1.25	320	6.67	195
Institutional	2023 H1	0.28	1,446	27.27	1,379
NBER	2023 H1	0.55	547	32.26	558
All sources	2023 H2	0.76	1,710	26.08	1,507
arXiv	2023 H2	1.18	340	5.91	220
Institutional	2023 H2	0.66	1,370	29.53	1,287
NBER	2023 H2	0.92	543	37.48	555
All sources	2024 H1	1.93	1,812	25.76	1,615
arXiv	2024 H1	2.17	368	5.69	246
Institutional	2024 H1	1.87	1,444	29.36	1,369
NBER	2024 H1	2.61	574	36.77	582
All sources	2024 H2	2.98	1,949	26.66	1,714
arXiv	2024 H2	3.58	419	7.50	280
Institutional	2024 H2	2.81	1,530	30.40	1,434
NBER	2024 H2	3.76	611	36.92	623
All sources	2025 H1	6.21	1,963	26.78	1,703
arXiv	2025 H1	7.86	509	6.94	317
Institutional	2025 H1	5.64	1,454	31.31	1,386
NBER	2025 H1	8.18	599	40.49	610
All sources	2025 H2	7.24	2,211	24.01	1,841
arXiv	2025 H2	6.99	730	7.54	451
Institutional	2025 H2	7.36	1,481	29.35	1,390
NBER	2025 H2	8.15	577	39.35	582
All sources	2026 H1	11.43	2,694	22.00	2,164
arXiv	2026 H1	13.21	1,037	6.53	597
Institutional	2026 H1	10.32	1,657	27.89	1,567
NBER	2026 H1	13.84	701	32.46	724
All sources	2026 H2*	21.08	2,196	21.01	1,666
arXiv	2026 H2*	24.58	1,139	5.58	699
Institutional	2026 H2*	17.31	1,057	32.16	967
NBER	2026 H2*	18.09	503	40.57	493

Timing-supported papers, with outcome-specific assessable denominators. Shares are percentages; counts are denominators. Institutional includes NBER. *July–October 2026 is incomplete.

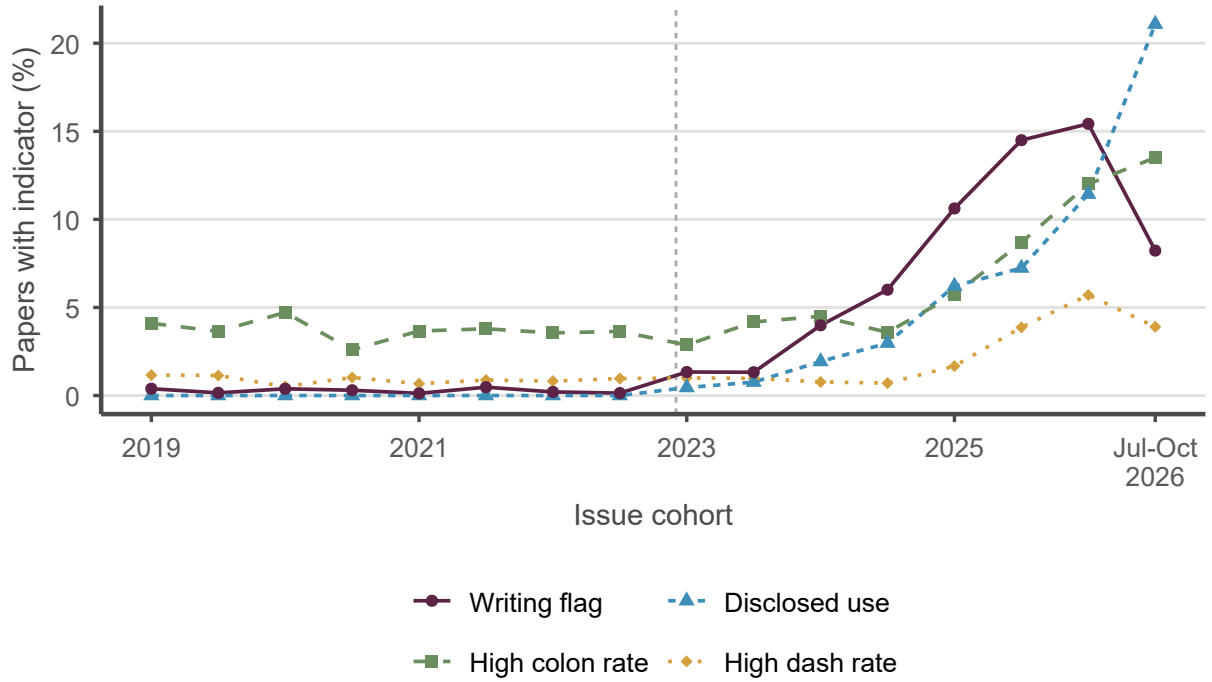


Figure A1. Trends in AI indicators

Half-year positive-indicator shares among timing-supported papers, using each measure's assessable denominator. The last observations cover July–October 2026, an incomplete half-year. High punctuation denotes rates above fixed historical thresholds. Issue dates need not date every word in a retrieved revision; punctuation includes some non-prose marks. Definitions appear in Appendix A.

A.2 Measurement and researcher identities

Assistance incidence follows a literal research-assistance role rule, with named counts requiring evidence identifying both person and role. Readers and advisers are recorded separately. Broader data, execution, and coordination support also enters the definition of solo authorship without assistance. Ambiguous role phrases remain uncertain rather than becoming negative observations. A collective credit can establish assistance while no individuals are named: zero named people does not imply zero total people. Incomplete named-count evidence remains missing. Negative incidence requires usable acknowledgement coverage; empty or unreadable extraction does not supply a zero.

The writing implementation pins `econ-ai-detector` commit `11285447a3cc` and version-3 neural revision `fb7fb8b27bb9`. A paper is flagged when its second-highest logistic window score is at least 0.33489 and its second-highest neural logit margin is at least 7.4827. The other released threshold pairs are $(0.18983, -3.7104)$ and $(0.14309, -5.4591)$. These are calibration choices, not guaranteed error rates in economics working papers. The continuous score is the second-highest logistic score, expressed in changes of 0.1 where used. Unsupported language, insufficient prose, and unresolved section selection remain missing.

Own-use disclosures require statements about generative AI's contribution to producing the pa-

per. Full-text tool and model terms identify candidate passages for contextual classification, which distinguishes own use, AI as a research topic, conventional tools, explicit non-use, and ambiguity. Positive evidence is checked against a literal source passage. Translation, writing, coding, search, data preparation, analysis, derivation, and checking labels can overlap. Under the whole-document task rule, a disclosure is writing-only when every supported task concerns manuscript drafting, editing, translation, or typesetting; research coding, data annotation, substantive search, estimation, derivation, or the generation of experimental inputs make it operational or mixed, and unspecified brainstorming or feedback is unclear rather than writing-only. Models that are studied, benchmarked, or supplied to experimental participants do not establish own use. Generic AI, or a grammar or translation service without evidence of generative functionality, remains unresolved. Saved prompts and per-call model identifiers accompany the replication code; Section A.4 describes the classification procedures and their model-provenance limits. This is measurement from text with explicit coding rules, not an inference of actual use from missing statements (Ludwig, Mullainathan and Rambachan, 2026).

Each confirmed disclosure passage was also coded for its content. Protective wording denotes an explicit restriction on AI's intellectual role or a claim that humans reviewed, checked, validated, or controlled the output; naming a task, saying that a tool was used sparingly, or accepting responsibility alone does not qualify. Numerical agreement between models and validation reproduced from cited studies do not count as human oversight. Separate labels record the human-oversight and limitation components, responsibility language, whether an individual operator is named, whether the wording is compatible with a publisher template, and whether reported writing use is confined to surface editing, proofreading, translation, or typesetting. The labels were proposed by the reviewing model from the retrieved contexts, every positive oversight or restriction proposal received a second source-context check, and the historical primary-session review adjudicated 34 cases against the rendered pages. The first-pass and adjudicated protective labels agree for 1,049 of the 1,167 disclosures. No human validation of these labels has been completed, and the checks share one manual, so shared systematic errors remain possible. Section B.5 reports the resulting comparisons.

Colons and em/en dashes are counted in the same selected material as the writing scores and expressed per 1,000 words. The pooled high-rate thresholds are 8.905 for colons and 9.267 for dashes, each fixed at the mean plus two standard deviations in a historical 2015–2021 reference sample. The raw dash measure includes numeric ranges and decorative rules; selected material can also retain tables and dash-shaped negative signs. The cleaned measure removes numeric ranges and repeated glyph runs. It remains a mixture of exposition, content, and typesetting. Source-specific historical flags and alternative constructions enter the complete supporting result family in Appendix B.3.

Researcher identities principally use OpenAlex works, identifiers, and corroborated author rosters. For recent papers, a flagged extension reuses an already linked identity where a full byline name is unique in the linked population. Ambiguous names remain unmatched. Table B8 excludes

these name-only extensions. Byline counts are measured independently of identity linkage. An indexed identity can contain attribution error, and the included linked researchers need not cover every printed author. Equation 2 gives each paper total weight one across its included researchers.

Additional outcomes distinguish comments/readership, solo authorship, and solo authorship without assistance, and transform named counts using $\log(1 + \text{count})$ or a cap at eight names, the fixed 99th percentile of the pre-2019 reference count distribution. Eight documented attributes describe formal theory, computational execution, existing-data use, original information acquisition, remote implementation, physical implementation, public data access, and restricted access. All eight must be resolved for their sum to be observed. Non-documentation does not establish absence. Appendix B.4 examines components whose evidence can overlap with disclosure coding.

A.3 Samples and overlap

Table A3. From observed measures to within-researcher samples

Outcome	Observed measures	Linked complete	Final papers	Disclosure switchers
Any assistant (pp)	11,506	9,715	7,243	535
Named assistants	11,505	9,714	7,242	535
Authors	13,713	11,516	8,502	633
Pages	13,713	11,516	8,502	633
Documented attributes	11,912	10,129	7,251	487

All outcomes begin with 16,721 timing-supported post-2022 papers. Observed measures require the outcome, both indicators and controls. Linked complete cases additionally require usable researcher identities. Final papers survive recurrence and fixed-effect pruning. Switchers count included researchers with both disclosure values on different papers; this is not the number of treated papers.

Table A4. Post-2022 paper characteristics

Measure	Papers	Mean	SD	Median
Authors	16,720	2.890	3.118	3.000
Any assistant	13,784	0.245	0.430	0.000
Named assistants	13,783	0.697	1.989	0.000
Pages	16,721	48.903	25.128	46.000
Writing flag	14,017	0.084	0.277	0.000
Disclosed use	16,301	0.072	0.258	0.000
Solo author	16,720	0.166	0.372	0.000
Solo, no assistance	13,543	0.111	0.315	0.000

One observation per timing-supported paper, January 2023–October 2026. Each measure uses its observed denominator before requiring linked recurring researchers. Binary measures are fractions. These samples differ from the regression samples.

The 13,713 post-2022 papers observed on both principal indicators contain 992 disclosures, 1,139 writing flags and 137 joint positives. These are jointly observed descriptive counts before the

recurrence and outcome restrictions used in the regressions. Of these papers, 11,719 carry neither recorded indicator, 855 disclose without a writing flag and 1,002 carry a writing flag without disclosure. The overlap percentages in the main text condition on the relevant positive-indicator denominator, not on all papers.

For a partition into cells c, the independence reference is $\sum_c n_{Dc} n_{Wc}/n_c$. Its expected joint count is 82.40 when pooled, 107.70 within source-year cells and 100.82 within source-quarter cells. The observed-to-expected ratios are respectively 1.66, 1.27 and 1.36. These are descriptions of recorded indicators on the same papers, not estimates of the correlation between unobserved uses. They allow for common variation in indicator prevalence without attributing low overlap to a single cause.

A.4 Measurement quality

Disclosure measurement combines candidate retrieval, document-level review, and literal source adjudication. The initial historical classification used GPT-5.6 Luna. A broader corpus-wide screen then retrieved passages for contextual review with GPT-6 Astra and tools disabled. Reviews distinguish author assistance from topical mentions, explicit non-use, and unresolved statements. A positive label requires a literal source passage; no statement in assessable text is coded negative, and missing or ambiguous evidence remains unresolved. Historical labels on retained documents are preserved. New-source review covers 4,785 source hashes and 1,079 document decisions, with positive oversight and restriction proposals checked against their surrounding passages. These procedures are model-assisted documentary measurement, not independent human validation.

The selected frame contains 1,729 positive disclosure labels, 53,617 zeros, and 1,239 missing values. Review and staging receipts support the requested GPT-6 Astra model and disabled tools. Component-review events do not independently identify the runtime model: their model attribution rests on the component protocol, unchanged command source, and coordinator record rather than a separate per-batch model receipt. This limits the provenance of the component classifications. Lexically selected candidate review does not establish corpus-wide recall. A stratified historical draw contains only 36 resolved reference labels among the 584 records that match the retained frame.

Task classifications apply to the whole document. A statement about editing does not make a paper writing-only if another passage reports AI-assisted research operations. Positive human-oversight labels require evidence of human checking; numerical agreement, validation reproduced from cited studies, and model-to-model agreement do not qualify. Missing statements and incomplete source text can still produce measurement error. Source evidence, decision ledgers, coding rules, and full outputs accompany the replication documentation.

A.5 Reference lists and institutional affiliation

This section reports the within-researcher reference-list and university-affiliation comparisons summarised in Section 5.5, alongside bibliography coverage and conditioning checks across all 24 series.

In the opening summary table, disclosing papers cite literature with a median age 1.24 years younger (SE 0.31; Romano–Wolf $p = .0005$; Holm $p = .0003$). Bibliography length differs by $+0.044$ $\log_2$ references (SE 0.041; Holm $p = .57$), and the three other tests in the four-test reference family all have Romano–Wolf p above .14. Among disclosing papers, 19.3% mention 'artificial intelligence' or 'AI' in the title or abstract against 2.5% of other papers; excluding them lowers the age gap to 0.54 years (SE 0.32). Adding paper-level university-tier shares leaves the disclosure estimates for any assistant and total pages unchanged at 6.63 percentage points and 5.41 pages (matched baselines 6.60 and 5.35).

Table A5. Reference lists and institutional composition

	Median age (years)	Reference count (log base 2)
Writing flag	-0.539	0.040
	(0.267)	(0.042)
Holm probability	0.131	0.565
Disclosed use	-1.239	0.044
	(0.312)	(0.041)
Holm probability	0.000	0.565
Papers	8,125	8,392
Disclosed papers	512	534
Researchers	5,084	5,218
Sources	24	24
University composition: disclosure coefficient		
	Any assistant (pp)	Pages
Matched baseline	6.599	5.346
	(2.606)	(1.274)
University tier shares	6.631	5.410
	(2.606)	(1.270)
Papers	7,243	8,502

Joint indicators and the specification in equation 2. Upper panel: counts use log base two of one plus the number of references. Eligible reference measures determine the corresponding fit; missing measures are withheld. Lower panel: paper-level shares of authors in THE 2026 university tiers, with unresolved slots explicit. Each baseline uses identical papers to its augmented comparison. Parentheses show researcher/paper-clustered standard errors. Source and shared-author sensitivities are reported separately.

Reference measurement covers all 24 series and uses the exact PDFs for the union of the ten fitted samples and the pre-2022 historical samples: 8,502 contemporary and 6,563 historical papers in the frozen measurement frame. Counts include distinct entries, including undated and non-journal items. Age uses the verified version year where available and the printed-date assumption otherwise. It requires at least five dated entries, dates for 80% of entries, and no unresolved future citation years. Incomplete lists and known extraction errors remain missing. Counts are available for 8,416 contemporary papers and ages for 8,198; Table A6 reports coverage by source and period.

Independent model-assisted text readings assess bibliography extraction in stratified documentary samples. Count agreement permits two entries or 5% of the independently enumerated count,

whichever is larger; age agreement requires the same eligibility decision and a median within one year. A historical 82-paper diagnostic sample gives count agreement of 21/24 for NBER, 4/4 for arXiv and 49/54 for other series; age agreement is 23/24, 4/4 and 52/54. A separate source-period-stratified sample gives weighted count/age agreement of NBER (8 papers; 100.0%/100.0%); arXiv (2 papers; 50.0%/100.0%); other series (14 papers; 85.7%/100.0%); Bank of England (30 papers; 97.3%/100.0%); Boston Fed (30 papers; 93.3%/95.8%). The samples cover distinct source-document batches. Their agreement rates are diagnostic and are not pooled into a population accuracy estimate.

Historical rendered-page checks include a 27-paper draw from series other than NBER and arXiv, with 98.1% weighted count agreement (25/27) and 94.9% age agreement (24/27). A separate draw completes 6 of 10 selected readings, with count agreement on 6/6 and age agreement on 6/6. Bibliographies exceeding the eight-page review limit are recorded as unavailable for this check. These fallible readings reveal merged entries, appendix material counted as references, and disputed citation years. Affected measures are withheld, without an accuracy threshold for admitting a series.

The additional reference audit gives weighted count agreement of 87.34% across 86 selected papers and age agreement of 95.14% across 85 eligible papers. Missing documentary readings remain in these denominators. The rendered-reference draw completes three of ten selected readings; all three agree on count and age, and seven are unavailable. These rates are reported separately from the historical samples. Known failed measures remain withheld.

Across the 24 series, median reference counts are 43 in the historical sample and 47 in the contemporary sample; median ages are 7 and 8 years, respectively. These are descriptions of selected papers, not a panel of unchanged projects. Regression counts use $\log_2(1 + \text{references})$ and age enters the human-input comparisons per five years. Table A7 retains identical-sample baselines. The four reference-outcome coefficients and fifteen reference-conditioned coefficients form separate Holm families. In the strictly dated sample, which contains NBER and arXiv papers, the disclosure coefficient for age is -0.77 years (standard error 0.61; 2,213 papers). The single-source disclosure/age estimates are -1.04 years (0.53; 2,854 papers) for NBER and -0.64 (0.73; 1,604 papers) for arXiv. Full single-source, timing, and shared-author sensitivities accompany the results.

Source-score tests give Holm-adjusted probabilities of 0.002 for disclosure in the age regression and 0.360 in the count regression, with about 3.6 effective sources for age. Shared-author covariance leaves their standard errors close to those in Table A5. Thus coverage across many series does not make the identifying information evenly distributed across them.

University affiliation comes from the measured paper's author-information pages, with same-year, own-DOI preprint metadata as a restricted fallback. Each author receives the best explicitly affiliated university's category in the THE overall 2026 ranking. This fixed edition contains 2,191 ranked institutions and 927 reporters. Categories are top 100, 101–200, 201–500, 501+, known unranked university, non-university, and unresolved; research-network membership does not substitute for an employer. This is a retrospective institutional classification, not a measure of an

Table A6. Reference coverage by source and period

	Historical			Contemporary		
Source	Target	Count	Age	Target	Count	Age
arXiv economics	534	521	481	1,872	1,847	1,775
Bank of Canada	106	105	102	60	58	59
BIS	99	97	92	95	92	93
Bank of England	77	76	74	54	52	51
Warwick CAGE	90	88	86	75	74	71
Cambridge	106	105	99	86	86	82
US Census CES	20	20	19	74	73	73
LSE CEP	117	115	112	72	72	72
CESifo	1,163	1,154	1,122	1,516	1,508	1,486
Osaka ISER	66	65	62	43	42	42
ANU CAMA	116	114	109	81	81	81
Boston Fed	22	19	20	35	31	31
San Francisco Fed	30	29	29	46	44	45
Federal Reserve Board	119	117	115	138	138	135
Chicago Fed	23	23	23	30	28	29
Institute for Fiscal Studies	46	46	45	85	84	84
NBER	2,724	2,701	2,613	3,123	3,101	3,018
New York Fed	55	53	50	45	45	44
Philadelphia Fed	62	61	59	60	59	56
RIETI	90	90	88	31	29	30
Tinbergen	102	101	99	70	70	69
Toulouse	120	115	111	158	157	153
Pompeu Fabra	46	44	40	9	9	8
World Bank	630	627	594	644	636	611

Each row is a source series. Historical and contemporary target counts are the frozen measurement-frame populations; count and age columns report the corresponding paper-level measures available for analysis. Missing measures are withheld, not imputed. Age requires at least five dated references, dates for at least 80% of entries and no unresolved future citation years.

economist's individual quality.

The mapping resolves 84.6% of 26,242 printed author slots. A historical 100-paper probability sample yields 76.4% weighted tier agreement against independent model-assisted front-matter readings and a separate ranking match (paper-clustered 95% interval 67.5–85.4%). A separate blinded rendered-page sample contains 18 papers and 56 author slots. Of these, 12 paper readings were completed; 6 could not be completed within the fixed measurement budget. Counting unavailable readings and unresolved decisions as disagreement gives 61.1% weighted tier agreement using the same frozen ranking crosswalk. These readings do not independently validate the crosswalk, and the diagnostic samples are not pooled. In the authorship sample, unresolved slots make up 16.4% of printed slots on disclosing papers and 15.4% on other papers, a difference of 1.0 percentage points that falls to 0.3 after source standardisation. The additional affiliation audit gives 61.7% weighted tier agreement across 105 slots in 28 selected papers. Readings were completed for 24

Table A7. Research inputs conditional on reference characteristics

	Any assistant (pp)	Named assistants	Authors	Pages
Writing: matched baseline	-3.459	-0.120	0.012	-0.768
	(2.327)	(0.117)	(0.063)	(1.152)
Writing: reference measures	-3.834	-0.127	-0.002	-0.743
	(2.344)	(0.119)	(0.063)	(1.149)
Disclosure: matched baseline	8.115	0.618	0.219	5.357
	(2.712)	(0.177)	(0.097)	(1.302)
Disclosure: reference measures	7.692	0.616	0.187	5.414
	(2.696)	(0.177)	(0.096)	(1.301)
Median age / 5	-1.447	0.014	-0.127	0.230
	(0.579)	(0.033)	(0.015)	(0.328)
Reference count	4.098	0.178	0.011	–
(log base 2)	(0.865)	(0.042)	(0.027)	
Papers	6,923	6,923	8,099	8,125
Disclosed papers	444	444	512	512

Each baseline uses exactly the rows, papers and paper-total weights of its corresponding augmented specification. Both AI indicators, original fixed effects and prose controls remain included. Researcher/paper-clustered standard errors are shown in parentheses; source and shared-author diagnostics accompany these estimates. Counts use log base two of one plus the number of references. Reference count is omitted for pages because bibliographies contribute directly to document length. The fifteen writing, disclosure, age and count coefficients form one Holm family.

papers; the four unavailable readings count as disagreements. These documentary checks do not establish human validation, and the samples are not pooled. Mapping errors limit institutional comparisons.

Tier shares divide by all printed authors, retaining unresolved slots and omitting non-university affiliation. Their author-count comparison conditions on its own denominator. Interactions use each row researcher's top-100 status and its identified main effect, excluding unknown statuses and recalculating recurrence and weights; all other resolved categories form the comparison group. The eight interactions have Holm-adjusted probabilities of at least 0.332, without establishing institutional equality. These models contain 6,129–7,045 papers. Each side has at least 195 disclosed papers from 13 sources and 176 researchers with within-status disclosure variation. Writing-flag support is at least 153 positive papers from 12 sources and 152 researchers with within-status variation. The eight tier-conditioned coefficients and eight interactions form separate Holm families. Table A8 reports estimates; complete-roster, NBER-omitted, and combined reference/rank comparisons accompany the results files.

Table A8. University affiliation and research inputs

	Any assistant (pp)	Named assistants	Authors	Pages
Writing: matched baseline	-3.206	-0.131	0.011	-0.693
	(2.251)	(0.111)	(0.061)	(1.100)
Writing: tier shares	-3.290	-0.134	0.004	-0.667
	(2.240)	(0.110)	(0.061)	(1.098)
Disclosure: matched baseline	6.599	0.548	0.240	5.346
	(2.606)	(0.168)	(0.093)	(1.274)
Disclosure: tier shares	6.631	0.551	0.233	5.410
	(2.606)	(0.168)	(0.093)	(1.270)
Writing × own top 100	-1.169	0.045	-0.133	4.808
	(4.854)	(0.264)	(0.138)	(2.358)
Disclosure × own top 100	-2.802	-0.124	0.137	0.901
	(5.144)	(0.317)	(0.191)	(2.514)
Papers: shares	7,243	7,242	8,502	8,502
Papers: interactions	6,130	6,129	7,045	7,045

THE overall 2026 ranking applied retrospectively to paper-specific affiliations. Tier shares divide by all printed authors, with unresolved slots explicit; their authors column is mechanically conditioned on the outcome denominator. Interaction coefficients use the row researcher's own paper-specific top-100 status, excluding unresolved statuses and recomputing recurrence and paper weights. Parentheses show researcher/paper-clustered standard errors. Eight AI coefficients with tier controls and eight institutional interactions form separate Holm families. These are descriptive institutional comparisons.

The disclosure-on-team-size comparison starts from the 8,502-paper authors sample. Raw disclosure shares for one, two, three, four, and five-or-more authors are 10.2%, 5.3%, 5.4%, 6.5% and 8.5%, respectively; denominators are 783, 2,200, 2,800, 1,735 and 984. The conditional log-base-two author-count slope is 1.36 percentage points (standard error 0.60). Disclosure information for that slope is concentrated in 4.1 effective sources, and its source-clustered interval spans zero. Giving eligible sources equal weight changes the slope from 1.21 to 2.17 percentage points on the same 9-source, 7,549-paper subset. On the 8,099-paper reference-and-rank subset, the matched slope is 1.18 percentage points (standard error 0.60), compared with 0.95 (0.60) after conditioning. These three slopes form one Holm family; none identifies author preferences.

B Supporting empirical results

B.1 Full outcomes and uncertainty

Table B2 gives all ten outcomes for the joint writing/disclosure model and the pooled comparison on identical samples. The two indicators across the ten within-researcher outcomes form a twenty-test family; Table B1 reports Holm-adjusted probabilities separately for researcher/paper clustering and source-score inference. Additional restrictions are descriptive sensitivities.

The source calculation is a null-restricted wild score test. Nuisance coefficients and fixed effects are estimated with the tested indicator's coefficient set to zero; source-level restricted scores receive Rademacher signs in 9,999 fixed-seed draws. It is an unstudentised score calculation, not a bootstrap-t interval. With disclosure information concentrated in about 3.3 effective sources, few influential signs can dominate its distribution. The calculation also does not automatically cover coauthor dependence across outlets, so small adjusted probabilities are conditional summaries rather than generally reliable staffing tests.

For shared authors, weighted, fixed-effect-partialled scores are first summed to each paper. The covariance sensitivity includes the outer products for the same paper and every pair sharing a reliably linked author, counting each ordered pair once even when several authors overlap. It allows covariance between whole papers, beyond rows bearing the same researcher label, but assumes no covariance outside linked neighbours. Incomplete linkage and broader source shocks can violate this assumption. Paper-only and overlap calculations omit small-sample corrections; the main researcher/paper standard errors use the software defaults.

In the assistance-incidence sample, linked identities cover 92.5% of byline slots. The largest connected component contains 60.8% of paper weight, and 1,844 linked authors span sources. Strict work-corroborated links cover 90.6% of slots, with 58.2% in the largest component. This extensive connection and incomplete coverage limit both one-hop covariance and any interpretation that assumes independent outlets. Table B3 reports the resulting sensitivities alongside the concentration of disclosure information.

Table B1. Family-adjusted probabilities for the complete outcome set

Outcome	Writing row clusters	Disclosure row clusters	Writing source score	Disclosure source score
Any assistant (pp)	1.000	0.170	1.000	1.000
Named assistants	1.000	0.019	1.000	0.036
Authors	1.000	0.163	1.000	1.000
Pages	1.000	< 0.001	1.000	0.028
Documented attributes	1.000	0.772	1.000	1.000
Solo author (pp)	1.000	1.000	1.000	1.000
Solo, no assistance (pp)	1.000	1.000	1.000	1.000
Any reader (pp)	1.000	1.000	1.000	1.000
log(1 + assistants)	1.000	0.004	1.000	0.537
Capped assistants	1.000	0.002	1.000	0.049

Holm adjustment across two indicators and ten outcomes, separately for each inference method. Row clusters means researcher/paper clustering. Source scores use 9,999 null-restricted Rademacher draws. Few effective sources and incomplete coverage of shared-author dependence prevent treating these probabilities as generally robust inference; a small value does not remove that limitation. These are exploratory comparisons.

Table B2. Pooled and within-researcher comparisons

Outcome	Writing pooled	Writing within	Disclosure pooled	Disclosure within	Papers	RW p W / D
Any assistant (pp)	-8.870	-3.206	8.949	6.599	7,243	.8185 / .157
	(2.142)	(2.251)	(2.422)	(2.606)		
Named assistants	-0.349	-0.131	0.585	0.548	7,242	.898 / .02
	(0.085)	(0.111)	(0.148)	(0.168)		
Authors	-0.010	0.011	0.207	0.240	8,502	.9975 / .147
	(0.072)	(0.061)	(0.100)	(0.093)		
Pages	-2.454	-0.693	4.973	5.346	8,502	.9935 / .0005
	(0.971)	(1.100)	(1.212)	(1.274)		
Documented attributes	-0.165	-0.083	0.249	0.133	7,251	.862 / .488
	(0.062)	(0.064)	(0.072)	(0.069)		
Solo author (pp)	3.319	1.331	3.994	0.359	8,502	.921 / .9975
	(1.712)	(1.282)	(2.302)	(1.658)		
Solo, no assistance (pp)	1.865	1.019	1.160	-0.144	7,099	.979 / .9975
	(1.463)	(1.295)	(1.606)	(1.373)		
Any reader (pp)	-4.254	-0.823	0.149	-1.178	7,136	.9975 / .9935
	(2.618)	(2.686)	(2.412)	(2.577)		
log(1 + assistants)	-0.121	-0.045	0.173	0.145	7,242	.7545 / .0035
	(0.026)	(0.029)	(0.039)	(0.039)		
Capped assistants	-0.306	-0.125	0.507	0.426	7,242	.6495 / .0015
	(0.065)	(0.075)	(0.111)	(0.110)		

RW reports Romano–Wolf probabilities for the within-researcher writing (W) and disclosure (D) tests, using 1,999 researcher-level wild-bootstrap draws and two-way studentisation across the twenty-test family. Each outcome uses identical papers in its pooled and within-researcher comparisons. Pooled specifications omit researcher effects; other controls and paper-total weighting follow equation 2. Standard errors in parentheses cluster by researcher and paper and are conditional on that structure. Binary outcomes use percentage points. All twenty within-researcher indicator/outcome tests enter the family in Table B1.

Table B3. Source concentration and shared-author uncertainty

Outcome	Effective sources	Largest share (%)	Paper SE	All-link overlap SE	Strict-link overlap SE
Any assistant (pp)	3.34	45.6	2.177	2.693	2.635
Named assistants	3.34	45.6	0.143	0.182	0.179
Authors	3.60	38.7	0.080	0.100	0.098
Pages	3.60	38.7	1.029	1.354	1.324

Effective sources is the inverse sum of squared source shares of residualised disclosure information. Overlap covariance sums whole-paper scores across each linked shared-author pair once. Strict links exclude name-only extensions. Paper-only and overlap calculations omit small-sample corrections; the main standard errors use the software corrections. Incomplete linkage and source shocks limit these sensitivities; they are not generally dependence-robust standard errors.

B.2 Documentation, source composition, and sample restrictions

The documentation comparisons use matched samples: each baseline retains exactly the papers in its corresponding conditioned specification (Table B4). Both include the same joint indicators and fixed effects. Full-text length conditioning is not reported for page count because it would condition document length on a closely related outcome.

Large-list and strict-timing restrictions appear in Table B5. Each refit re-evaluates eligibility and fixed-effect support and recomputes paper-total weights. These restrictions change the population being described. In particular, positive point estimates in the small strict-date sample do not establish temporal comparability of the larger printed-date sample.

The equal-source comparison retains sources with at least five disclosed and twenty no-statement papers in the final outcome-specific sample. Its ordinary-weight and equal-weight rows use identical papers (Table B6). For assistant incidence, the eight sources are arXiv economics, Warwick CAGE, Cambridge, CESifo, Federal Reserve Board, Institute for Fiscal Studies, NBER, and World Bank. Renewed recurrence and fixed-effect restrictions leave 6,387 papers for both weighting schemes. Equal weighting gives each source the same total weight, not the same residual information. The equal-source incidence interval spans negative and positive differences.

Table B7 reports the two largest-source omissions and the range over every single-source omission. Tables B6 and B8 distinguish source reweighting from stricter identity evidence. Complete source-by-outcome support, omissions, uncertainty, and weights accompany the result files; the ranges are not selected confidence bounds.

Table B4. Disclosure and documentation on matched samples

	Any assistant (pp)	Named assistants	Authors	Pages
Acknowledgement baseline	6.599	0.548	0.195	5.408
	(2.606)	(0.168)	(0.097)	(1.359)
Papers	7,243	7,242	7,718	7,718
Acknowledgement length	4.968	0.468	0.176	4.478
	(2.573)	(0.166)	(0.097)	(1.372)
Papers	7,243	7,242	7,718	7,718
Full-text baseline	6.599	0.548	0.240	–
	(2.606)	(0.168)	(0.093)	
Papers	7,243	7,242	8,502	–
Full-text length	5.742	0.485	0.218	–
	(2.593)	(0.167)	(0.094)	
Papers	7,243	7,242	8,502	–

Disclosure coefficients and researcher/paper-clustered standard errors. Each conditioned row is compared with its own identical-sample baseline. Acknowledgement length uses log(1 + words); full-text length uses log(words). Acknowledgement length is partly generated by the credits and disclosures themselves; these are conditional descriptions, not bias-corrected staffing effects. Source and author-dependence limitations apply.

Table B5. Large lists and document timing

	Any assistant (pp)	Named assistants	Authors	Pages
At most eight assistants	6.183	0.317	0.213	4.562
	(2.702)	(0.099)	(0.106)	(1.428)
Papers	7,131	7,131	7,131	7,131
At most ten authors	6.658	0.533	0.165	5.476
	(2.631)	(0.170)	(0.069)	(1.275)
Papers	7,210	7,209	8,469	8,469
Independently dated versions	-0.840	0.325	0.542	2.173
	(4.258)	(0.311)	(0.191)	(1.869)
Papers	1,615	1,615	2,368	2,368

Disclosure coefficients; standard errors in parentheses cluster by researcher and paper. Eligibility, fixed-effect support and paper-total weights are recomputed after each restriction. The unconditioned main estimates are in Table 1. Strict document timing concerns the dated text, not simply a file hash. Smaller samples and dependence limit inference.

Table B6. Ordinary and equal weights on identical source support

	Any assistant (pp)	Named assistants	Authors	Pages
Ordinary paper weights	6.544	0.511	0.217	5.146
	(2.665)	(0.173)	(0.096)	(1.323)
Papers	6,387	6,387	7,549	7,549
Equal source weights	1.977	0.361	0.296	8.865
	(4.667)	(0.168)	(0.108)	(2.456)
Papers	6,387	6,387	7,549	7,549

Disclosure coefficients and conditional researcher/paper-clustered standard errors. Each outcome retains sources with at least five disclosed and twenty no-statement papers, after renewed fixed-effect pruning. Equal weighting gives each retained source total weight one, divided equally across its papers and then linked-researcher rows. The matched ordinary-weight row isolates reweighting from the support restriction. Source concentration and shared-author dependence remain relevant.

Table B7. Dependence of disclosure estimates on source composition

Outcome	Omit NBER	Omit arXiv	Range across all omissions	Sources omitted
Any assistant (pp)	3.463	10.236	3.463 to 10.236	24
	(3.358)	(3.178)		
Named assistants	0.438	0.726	0.438 to 0.726	24
	(0.156)	(0.204)		
Authors	0.276	0.079	0.079 to 0.338	24
	(0.163)	(0.105)		
Pages	4.938	6.005	4.783 to 6.005	24
	(1.687)	(1.601)		

Each column refits the within-researcher model after a source omission, with eligibility and paper-total weights recomputed. Parentheses contain conditional researcher/paper-clustered standard errors. Ranges describe point estimates, not confidence bounds. The complete omission estimates, standard errors and sample sizes accompany the results files.

Table B8. Work-corroborated researcher identities

	Any assistant (pp)	Named assistants	Authors	Pages
Disclosure	7.655	0.614	0.201	5.499
	(2.741)	(0.178)	(0.097)	(1.335)
Papers	6,953	6,952	8,135	8,135

Excludes name-only link extensions and refits the within-researcher comparison. Authorship itself comes from the byline. Parentheses contain conditional researcher/paper-clustered standard errors. Corroboration narrows identity uncertainty but does not establish complete author linkage.

B.3 Punctuation and historical comparisons

Punctuation offers alternative measurements of text characteristics, with sensitivity to both construction and period. The contemporary family contains 380 distinct coefficient tests across original and cleaned rates, pooled and source-specific historical thresholds, raw-count flags, joint controls, alternative caps, and companion disclosure coefficients. Numerically identical tests are recorded as aliases, not independent evidence. Table B9 summarises rates under adjustment across that complete family. Continuous coefficients contrast one additional glyph per 1,000 selected words; threshold flags instead compare binary groups. The complete result files preserve those units and all outcomes.

The cleaned-dash coefficient for solo authorship has a full-family adjusted probability of 0.303. It therefore does not establish a smaller-team pattern after accounting for construction choice. Two colon/page comparisons retain adjusted probabilities below 0.001, namely the rate shown in Table B9 and the same rate in the specification that adds the writing flag and dash rate, but they concern text and document length rather than a verified AI effect. Shared-author and source-dependence qualifications apply to these exploratory comparisons too.

Historical specifications include each text measure and its interaction with the post-2022 period, with researcher, source-quarter, and field effects. Their eighty tests include levels and interactions for four constructions and ten outcomes. The post-2022 interactions supply no adjusted evidence of an AI-specific change. Pre-period associations also demonstrate that text characteristics can relate to research inputs before widespread generative AI. Selection, document timing, and exposition provide competing interpretations; a contemporaneous coefficient does not by itself distinguish them.

Table B9. Text-rate associations under the complete construction family

	Any assistant (pp)	Authors	Pages	Solo author (pp)
Colon rate	-0.171	0.003	-0.496	0.317
Full-family Holm	1.000	1.000	< 0.001	1.000
Dash rate	0.201	-0.007	0.074	0.340
Full-family Holm	1.000	1.000	1.000	1.000
Cleaned dash rate	0.201	-0.022	0.182	0.649
Full-family Holm	1.000	1.000	1.000	0.303

Each rate coefficient contrasts one additional glyph per 1,000 selected words. Joint disclosure controls and the within-researcher design remain included. Holm values cover 380 distinct contemporary coefficient tests, including construction alternatives and companion indicators. Complete pointwise intervals, standard errors and aliases are supplied in the results files. Rates include content/typesetting as well as prose; no AI-specific staffing effect is inferred.

B.4 Documented attributes

An operational disclosure can also establish that a paper uses a computational or data-related method. The full attribute count and the task classification can therefore share evidence. Table B10 removes formal theory, computational execution, and existing-data use, retaining the five acquisition, implementation, and access components. The remaining information is still not guaranteed independent: original data may, for example, include AI annotation. These counts describe documented attributes rather than latent breadth, and neither the full nor reduced comparison validates scope expansion.

Table B10. Documented attributes and shared evidence

Index	Indicator	Estimate	SE	Papers
Eight attributes	Writing	-0.083	0.064	7,251
Eight attributes	Disclosure	0.133	0.069	7,251
Five attributes	Writing	0.002	0.036	7,251
Five attributes	Disclosure	0.167	0.047	7,251

The reduced index omits formal theory, computational execution and existing-data use, retaining acquisition, implementation and access attributes. Both indices use identical papers. Standard errors condition on researcher/paper clustering. The reduced comparison is a sensitivity, not an independent measure of scope; original information acquisition may itself use AI-generated annotations. Its pointwise probabilities are not additional confirmatory tests.

Full estimates, pointwise uncertainty, adjustment families, construction definitions, and sample membership are retained with the replication code and results, including supporting specifications not displayed here.

B.5 Disclosure wording

This section reports the comparison of protective disclosure wording – explicit limits on AI's intellectual role or claims of human oversight – summarised in Section 4.2, alongside its sensitivity tables. Across all 1,145 fitted disclosures with source-year and field fixed effects and task and template controls, coauthored papers exceed solo papers in protective wording by 4.9 percentage points (SE 2.0; raw $p = .022$). That overall contrast fails the registered eight-test Holm adjustment ($p = .176$), which is retained here and in Appendix B.6 because 1.5 to 2.7 effective sources make a source-level bootstrap unreliable. Among the 162 fitted writing-only disclosures, the coauthored-minus-solo difference is 10.2 percentage points (SE 6.3; raw $p = .143$; wild source-cluster bootstrap $p = .27$). ArXiv supplies 80.3% of the coefficient score mass in that writing-only baseline, and omitting 2026 moves the estimate to -0.8 percentage points.

Protective wording is observed for 1,163 of the 1,167 recorded disclosures; four passages are unresolved. Shares by author count are 50.4%, 54.4%, 58.1%, 49.0% and 56.6% for one to five-or-more authors. The registered comparison (T06 in Table B14) fits a paper-level linear probability model on the 1,145 disclosures retained after fixed-effect pruning, with source-year and field effects,

Table B11. Protective wording in AI disclosures: coauthored minus solo papers

	Team minus solo (pp)	Papers	Solo papers	Sources
All disclosures	4.9 (2.0)	1,145	–	20
Writing-only disclosures	10.2 (6.3)	162	34	10
Human-oversight component	11.1 (5.0)	162	34	10
Limitation component	-7.7 (7.7)	162	34	10
Named tool or service in the statement	11.4 (5.0)	133	29	8
No template-compatible wording	18.1 (8.7)	102	26	9
Narrow editing tasks only	9.7 (7.8)	106	17	8
Statement-length control	13.3 (7.0)	162	34	10
Omitting 2026	-0.8 (9.7)	61	12	6

Linear probability models of protective wording among papers with a recorded own-use disclosure. The first row uses all fitted disclosures with source-year and field effects and task and template controls; the remaining rows use writing-only disclosures with source-year and field effects and a template indicator, except where the row removes template-compatible wording. Coefficients are coauthored minus solo in percentage points; source-clustered standard errors in parentheses. The human-oversight and limitation components are the two parts of the composite outcome. A null-restricted wild source-cluster bootstrap over all 1,024 sign patterns gives $p = 0.27$ for the writing-only baseline and $p = 0.25$ for the narrow-task row. ArXiv supplies 80.3% of the residual identifying variation in the baseline (1.53 effective sources); the shared-author covariance diagnostic gives a standard error of 9.8 points. Solo counts for the all-disclosure row are not separately recorded after pruning. These are exploratory conditional associations among disclosing papers, not confirmatory tests; Appendix B.5 reports every specification.

task controls, and a template indicator, clustered by source. Its solo-minus-coauthored contrast is -4.9 points with an ordinary interval from -9.0 to -0.8 and a Bonferroni interval from -11.0 to 1.1. Without the task controls the contrast is -9.3 points; among operational disclosures alone it is -3.0.

The writing-only comparison was a registered secondary split. Under the whole-document task rule, 175 disclosures are writing-only, 119 of them confined to surface editing, proofreading, translation, or typesetting, and all have a known outcome: 103 of 139 coauthored papers (74.1%) and 25 of 36 solo papers (69.4%) contain protective language. The estimation sample after singleton pruning has 162 papers from 10 sources: 96 of 128 coauthored (75.0%) and 23 of 34 solo (67.6%). Tables B12 and B13 report every specification in the follow-up protocol, which was frozen after the secondary estimate was first known and before further coefficients were inspected; the protocol also fixes the restriction to statements naming a tool or service before its estimation. All probabilities are unadjusted for the selection of this secondary comparison. Template-compatible wording means lexical compatibility with the review-and-responsibility formulation in a current publisher policy, not proven copying or a binding rule. The statement-length control uses the log word count of the extracted own-use clauses, which can itself reflect the reporting decision.

Inference is limited by source concentration. ArXiv supplies 80.3% of the coefficient score mass in the baseline, giving 1.53 effective sources. Every single-source omission retains the sign; excluding arXiv leaves 63 papers with six solo papers and a very wide interval. The restricted wild

source-cluster bootstrap enumerates all 1,024 sign patterns and gives $p = 0.27$; it is a symmetric-error approximation, not randomisation inference. A shared-author adjacency covariance gives a standard error of 9.8 points against 6.3 under conventional clustering, with 74.7% of author slots linked. Omitting 2026 leaves 61 papers and a difference of -0.8 points. A formal interaction between solo authorship and writing-only task on 1,061 disclosures gives -0.9 points (conventional $p = 0.88$, wild $p = 0.89$); allowing source-year and field effects to differ by task group gives 7.1 points ($p = 0.24$, wild $p = 0.21$). Raw protective shares among coauthored writing-only papers are 30 of 41, 40 of 50, 17 of 23 and 16 of 25 for two, three, four, and at least five authors, so the evidence concerns the solo-team contrast rather than a gradient.

The reviewing model's first-pass protective label, before source-context adjudication, gives a team-minus-solo difference of 9.4 points on the writing-only sample, compared with 10.2 after adjudication (Table B13). This is a stability check within one measurement procedure, not independent validation. The main author-count sample overlap has only four sources; its point estimate is reported without an interval or probability.

Table B12. Protective wording among writing-only disclosures: specifications and omissions

Specification	Team minus solo (pp)	95% interval	Papers	Solo	Sources
Focused specifications					
Baseline (source-year and field effects, template indicator)	10.2 (6.3)	[-4.1, 24.4]	162	34	10
Named tool or service in the own-use statement	11.4 (5.0)	[-0.3, 23.2]	133	29	8
Excluding template-compatible wording	18.1 (8.7)	[-2.0, 38.2]	102	26	9
Excluding repeated normalised wording	10.2 (6.3)	[-4.1, 24.4]	162	34	10
Source-year effects only	8.1 (6.5)	[-6.6, 22.9]	163	34	10
Source, year and field effects	3.3 (7.1)	[-12.4, 19.0]	170	35	12
Source-year cells with both solo and coauthored papers	11.3 (6.5)	[-4.0, 26.5]	124	34	8
Baseline rows without template control	16.4 (5.3)	[4.3, 28.4]	162	34	10
Baseline rows with statement-length control	13.3 (7.0)	[-2.6, 29.1]	162	34	10
Independently dated documents	8.2 (n/a)	unavailable	95	27	2
Unknown outcome coded 0	10.2 (6.3)	[-4.1, 24.4]	162	34	10
Unknown outcome coded 1	10.2 (6.3)	[-4.1, 24.4]	162	34	10
Period and source omissions					
Omitting 2024	11.2 (6.6)	[-3.7, 26.1]	148	33	10
Omitting 2025	14.7 (8.4)	[-4.7, 34.0]	110	22	9
Omitting 2026	-0.8 (9.7)	[-25.7, 24.1]	61	12	6
Range over the 10 single-source omissions	7.4 to 30.4				

Coefficients are coauthored minus solo in percentage points with conventional source-clustered standard errors and intervals; the sign of the fitted solo coefficient is reversed. The baseline has source-year and field effects and a template indicator. The repeated-wording row applies the frozen normalisation rule. Source-cluster intervals are withheld with fewer than five sources. All probabilities are exploratory and unadjusted for selection of this secondary comparison.

B.6 Registered mechanism comparisons

The register of eight comparisons was frozen before any of their coefficients was estimated, after a feasibility stage that reproduced the main estimates to numerical precision and inventoried cov-

Table B13. Protective wording among writing-only disclosures: components, interactions, label stability and sample overlap

Specification	Team minus solo (pp)	95% interval	Papers	Solo	Sources
Outcome components (baseline sample)					
Human-oversight component	11.1 (5.0)	[-0.2, 22.4]	162	34	10
Limitation component	-7.7 (7.7)	[-25.2, 9.8]	162	34	10
Responsibility component	-7.3 (6.5)	[-22.1, 7.5]	162	34	10
Task interactions (writing-only and operational/mixed disclosures)					
Writing-only minus operational/mixed, task controls	-0.9 (6.2)	[-14.0, 12.1]	1,061	192	20
Same, excluding templates	-2.8 (8.1)	[-19.8, 14.3]	966	173	20
Writing-only minus pure operational	-6.4 (7.3)	[-21.9, 9.1]	774	98	18
Same, effects varying by task group	7.1 (5.9)	[-5.2, 19.5]	1,051	190	19
Label stability and narrow editing tasks					
First-pass label before source adjudication, baseline sample	9.4 (6.1)	[-4.4, 23.3]	162	34	10
Narrow editing tasks only, protective wording	9.7 (7.8)	[-8.8, 28.1]	106	17	8
Narrow editing tasks only, human-oversight component	5.0 (5.3)	[-7.6, 17.5]	106	17	8
Narrow editing tasks only, limitation component	2.7 (10.4)	[-21.9, 27.3]	106	17	8
Overlap with the manuscript author sample					
Papers in the main author-count sample	-19.8 (n/a)	unavailable	64	8	4
Papers with complete linked rosters	15.8 (1.3)	[12.1, 19.5]	98	27	5
Papers with at least one linked author	12.2 (5.3)	[-1.5, 25.9]	128	28	6

Same conventions as Table B12. Task interactions compare the solo association in writing-only disclosures with operational or mixed-use disclosures; their papers and solo columns count the combined sample. Wild source-cluster probabilities use a null-restricted residual bootstrap with Rademacher signs: 9,999 draws give 0.89 and 0.21 for the first and fourth interaction rows, and enumeration over all sign patterns gives 0.25 for the narrow-task protective row. The first-pass label is the reviewing model's proposal before the source-context adjudication of positive oversight and restriction claims; the comparison is a stability check within one measurement procedure, not independent validation. Sample-overlap rows restrict to papers linked to the manuscript's researcher panel; linking is selective and these are not independent replications.

erage. The data had already been analysed, so the freeze is an inductive-analysis protocol, not a preregistration. The registered specifications, contrasts, and family are retained with the adjudicated disclosure labels. All tests except T06 reverse the outcome of the main regressions: they ask which predetermined team characteristics predict a recorded disclosure, in weighted researcher-paper linear probability models with researcher, source-quarter, and field effects, author-count bins, and the prose controls, each paper carrying total weight one. They do not estimate a share of the author-count coefficient explained. Holm adjustment applies across the eight entries, with unestimable entries carried at $p = 1$. Table B14 and Figure B1 report the six estimable entries with ordinary and Bonferroni intervals, adjusted probabilities, every single-source omission, and the concentration of residual information.

Historical features require complete linked rosters and observed pre-2022 histories; missing histories are not assigned zeros. Complete rosters cover 794 of 1,167 disclosing and 10,508 of 15,133 other papers, and complete pre-2022 standing covers 205 and 4,701 of them. The field-overlap, standing, and standing-range models therefore describe a selected population of 2,401 papers with 87 disclosures. Standing is a citation percentile within the linked historical population, not a global

Table B14. Registered comparisons of explanations for disclosure and team size

Test	Feature	Papers / disclosures	Estimate (pp)	Ordinary 95% interval	Bonferroni interval	Holm p	Effective sources
T01	Field overlap	2,401 / 87	-0.95	[-3.07, 1.17]	[-3.90, 2.01]	1.000	2.51
T02	Mean historical standing	2,401 / 87	1.03	[-1.40, 3.45]	[-2.36, 4.41]	1.000	2.56
T03	Standing range	2,401 / 87	-0.37	[-2.83, 2.10]	[-3.80, 3.07]	1.000	2.54
T05	Coauthor documentation	547 / 14	2.02	[-4.30, 8.34]	[-6.82, 10.87]	1.000	1.72
T06	Solo protective wording	1,145 / 1,145	-4.91	[-9.02, -0.79]	[-10.95, 1.14]	0.176	2.05
T08	Submission-stage signals	8,502 / 544	7.15	[-0.94, 15.23]	[-4.13, 18.43]	0.582	2.72
T04	Prior coauthor propensities	Unestimable: one paper has complete prior outside-team histories					
T07	Detection recall	Unidentified: solo minus multi-author recall bounds $[-0.96, 0.99]$					

Eight entries frozen before estimation and re-estimated unchanged on the reviewed disclosure labels; Holm adjustment across all eight with unestimable entries carried at $p = 1$; Bonferroni intervals are simultaneous across the eight. Figure B1 shows every single-source omission. All entries except T06 use recorded disclosure as the outcome in weighted researcher-paper linear probability models with researcher, source-quarter and field effects, author-count bins and the prose controls of equation 2; T01 adds mean standing and career age, T02 and T03 enter mean and range jointly with age. T06 uses protective wording among disclosures with source-year and field effects and task and template controls, clustered by source; other intervals cluster by researcher and paper. Contrasts: T01, pairwise agreement in authors' pre-2022 broad fields, 0 to 1; T02, interquartile change, 24.7 percentile points; T03, interquartile change, 41.6 percentile points; T05, interquartile change in coauthors' pre-2022 log acknowledgement length; T06, solo minus coauthored, among disclosures; T08, presence of venue notice or data and conflict declarations. Source-omission estimates and any explicit failures are retained in the accompanying CSVs. Every estimable entry fails the prespecified source-support screen; effective sources are the inverse sum of squared source shares of residual information.

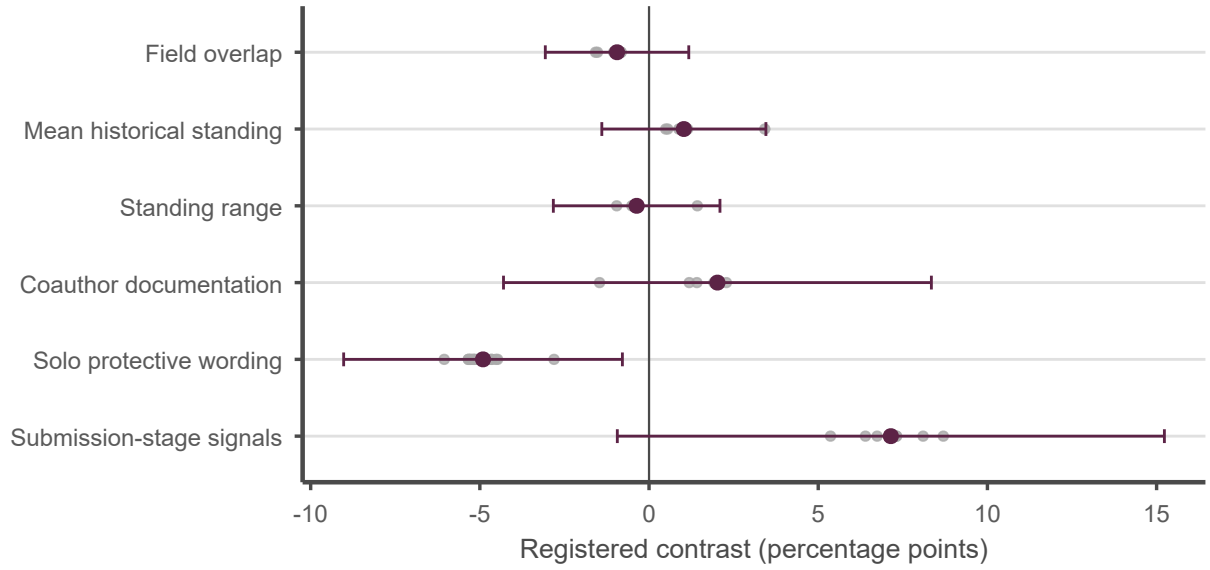


Figure B1. Registered contrasts and single-source omissions

Plum points and bars: registered contrast and ordinary 95% interval from Table B14. Grey points: the contrast after omitting each source in turn, with eligibility and weights recomputed. T05 without NBER is estimable and is included. Continuous features use the frozen contrast units. No entry survives Holm adjustment, and every entry fails the source-support screen.

rank, and field overlap is pairwise agreement in dominant broad fields, a weak proxy for what a reader infers about the operator. Coauthor documentation uses the mean of coauthors' pre-2022 log acknowledgement length; its 14 disclosures leave the estimate dependent on NBER, but omitting NBER remains estimable, with a registered contrast of -1.46 percentage points. Prior coauthor disclosure propensities require at least three independently dated earlier papers outside the current team for every coauthor, which one paper satisfies.

The submission-stage indicator is an explicit venue submission or acceptance notice, or the combination of a structured data-availability statement and a conflict-of-interest declaration; the AI passage itself is excluded. Only 98 papers in the 8,502-paper sample carry it, of which 16 disclose, all through paired declarations without an explicit notice; the 17 papers with explicit notices contain no recorded disclosure. On the 547-paper sample with complete coauthor documentation histories, the stage contrast is 1.3 points before and 1.12 after adding documentation, with 14 disclosures. On the 7,718-paper sample with acknowledgement text, it is 12.7 points before and 13.3 after conditioning on acknowledgement length, and -6.4 and -5.9 after removing verified AI clauses from the acknowledgements on a 1,485-paper sample. The coefficients also depend on the sample and on removing the AI clauses; pooling the main rows without researcher effects gives 6.1 points. Secondary comparison outcomes show disclosure associated with 0.49 more advice or comment credits (interval -1.62 to 2.59) and -0.2 points in literal software mentions (-0.9 to 0.5); neither is a guaranteed negative control.

Differential detection cannot be identified from the existing audit. Of its 825 documents, 584 match the current frame and 36 have resolved reference labels. Design-weighted bounds that let unresolved recorded negatives contain missed statements and unresolved positives be false give a solo recall interval of $[0.04, 1.00]$ and a multi-author interval of $[0.01, 1.00]$, so the difference lies in $[-0.96, 0.99]$. A 20-document whole-text pilot covering 1,255 pages left four originally negative documents unresolvable because of garbled mathematics and image-only material, so the planned 300-document text-only audit was not run; a valid reference standard requires page-image adjudication. The second review of candidates described in Section A.4 does not supply such a standard, because its candidates were selected by a lexical screen.

The opportunity benchmark is the complementary log-log regression of the disclosure indicator on $\log n$ with source-year cell effects, which the independent constant-probability model restricts to a unit coefficient. The free coefficient is 0.116 (standard error 0.077) on 14,476 papers in 60 cells with outcome variation; 1,824 papers in constant cells are excluded, and overlapping authors are not fully addressed by its covariance. The conditional author-count curve in Figure 2 uses the 8,502-paper sample; its all-author slope of 1.36 points per doubling has a source-clustered interval from -0.67 to 3.38, and the coauthored-only slope of 2.26 points on a re-pruned 7,599-paper sample has a source-clustered interval from -1.08 to 5.60. Descriptive standardisation on common source-year-field support leaves solo and pair rates of 8.9% and 7.3%; adding historical career-age support gives 6.6% and 3.8% on a smaller population. Among solo papers with assessable credits, disclosure is 10.5% without a named assistant credit and 6.1% with one.

B.7 Assistant-credit portfolios

Credited assistance can be reused across a researcher's projects, so credits per paper need not track distinct people. For each linked researcher and calendar year, the portfolio comprises their papers Q in the 22 sources represented in every year from 2015 to 2025, the credited assistant-paper links L, and the distinct assistants H, with credits per paper $a = L/Q$ and papers per assistant $r = L/H$, so that $H = Qa/r$. Complete acknowledgement coverage is required; 98,577 of 107,836 candidate portfolios are retained across the full-year and January–September windows, which overlap. Cross-paper assistant identities lack corroborating institution and date evidence, so H is reported as identity-free bounds between the largest single-paper count and the total number of credits, with exact-name merging shown as an unverified scenario. These portfolios do not depend on disclosure measurement.

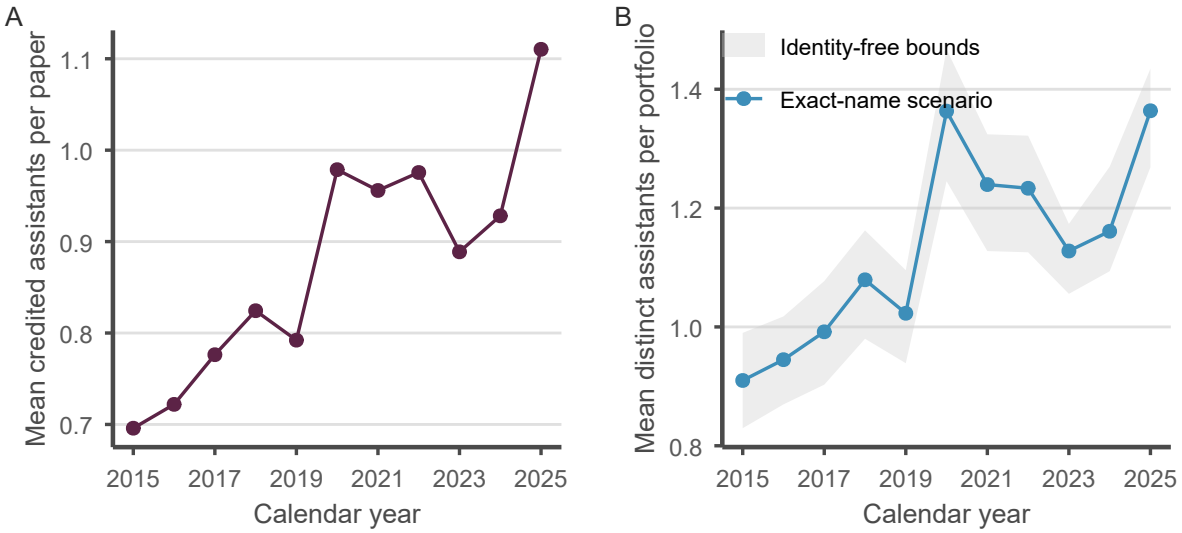


Figure B2. Credited assistants within annual researcher portfolios

Full-year portfolios, 2015–2025, of linked researchers with complete acknowledgement coverage in sources present in every year. Panel A: mean credited assistant-paper links per paper. Panel B: mean distinct assistants per portfolio under the unverified exact-name scenario, with identity-free bounds in grey. Researcher composition changes across years, portfolios of coauthors overlap and cannot be summed as a workforce, and counts refer to credits rather than employment or hours.

Mean credits per paper rise from 0.98 in 2022 to 1.11 in 2025, with a dip to 0.89 in 2023. For 2025 the identity-free bounds on distinct assistants per portfolio are 1.27 to 1.43. The January–September 2026 window is compared only with the same months of earlier years. These descriptions do not establish increased reuse, employment displacement, or an AI effect; the cross-paper reuse rate remains unresolved until assistant identities are corroborated.

B.8 Exploratory decompositions and step-down inference

This section collects the supporting tables and figure for the analyses added after the registered protocol was frozen, reporting raw p-values for every exploratory outcome and indicator split. Task-mix regressions appear first, pairing descriptive means with pooled and within-researcher estimates for operational-or-mixed and writing-only disclosures outside the registered twenty-test family. Page-split diagnostics follow across 8,493 verified texts, reporting 60-paper boundary-audit error rates alongside single-source omissions and estimates on the 1,225 bookmark-confirmed papers. Multiple-testing results and Figure B3 cover 41,828 papers, pairing 160-snippet precision checks and half-year shares since 2015 with post-2022 source-year and 8,193-paper within-researcher models. Last, the step-down inference table sets 1,999-draw researcher-bootstrap Romano–Wolf p-values beside the registered Holm values across all twenty main tests and the four reference-list tests.

Table B15. Task mix: pooled comparisons

Outcome	Operational	Writing-only	Equality p
Any assistant (pp)	11.149 (2.591)	6.118 (4.959)	.3558
	$< .0001$	.2173	
Named assistants	0.753 (0.173)	0.160 (0.214)	.0233
	$< .0001$	.4539	
Authors	0.314 (0.104)	0.345 (0.177)	.8747
	.0026	.051	
Total pages	7.113 (1.384)	0.571 (2.225)	.0108
	$< .0001$	.7975	

Pooled regressions omit researcher effects but retain source-quarter and field effects, the writing flag and prose controls. Entries show coefficients, researcher/paper-clustered standard errors and raw p-values; the last column tests equality. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B16. Task mix: descriptive means

Outcome	Disclosure group	Papers	Mean
Any assistant (pp)	no-disclosure	6,775	28.753
Any assistant (pp)	operational-or-mixed	361	37.119
Any assistant (pp)	writing-only	68	23.529
Any assistant (pp)	unlabelled	39	15.385
Named assistants	no-disclosure	6,774	0.829
Named assistants	operational-or-mixed	361	1.518
Named assistants	writing-only	68	0.647
Named assistants	unlabelled	39	0.385
Authors	no-disclosure	7,958	3.078
Authors	operational-or-mixed	425	3.327
Authors	writing-only	78	3.295
Authors	unlabelled	41	1.902
Total pages	no-disclosure	7,958	52.430
Total pages	operational-or-mixed	425	62.259
Total pages	writing-only	78	49.654
Total pages	unlabelled	41	43.439

Paper-level, unadjusted means by disclosure group. Assistant incidence uses percentage points; other outcomes use counts or pages. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B17. Task mix: support and within-researcher variation

Outcome	O	W	U	O switchers	W switchers	Both
Any assistant (pp)	361	68	39	441	90	13
Named assistants	361	68	39	441	90	13
Authors	425	78	41	521	111	17
Total pages	425	78	41	521	111	17

O, W and U count operational-or-mixed, writing-only and unlabelled disclosed papers. Switchers are researchers with within-portfolio variation in the indicated dummy. Both counts researchers observing both O and W papers. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B18. Page boundaries: restricted-sample disclosure estimates

Sample	Outcome	Estimate (SE), p	Papers	Disclosed
Omit NBER	Main-text pages	0.26 (0.92)	5,131	343
		.7752		
Omit NBER	Appendix pages	4.53 (1.47)	5,131	343
		.002		
Omit arXiv	Main-text pages	2.48 (0.83)	6,553	341
		.0027		
Omit arXiv	Appendix pages	3.41 (1.40)	6,553	341
		.0149		
Bookmark-confirmed	Main-text pages	4.07 (1.89)	1,225	48
		.0313		
Bookmark-confirmed	Appendix pages	6.23 (3.57)	1,225	48
		.0813		

Frozen pages-sample weights, original controls and fixed effects; researcher/paper-clustered standard errors. Raw p-values accompany each coefficient. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B19. Held-out page-boundary validation

Boundary	Reviewed	Errors	Error rate (%)
Appendix boundary	60	7	11.7
References boundary	60	4	6.7

Held-out documentary checks on papers not used for tuning. Unweighted rates; weighting by prevalence gives an appendix-boundary error rate of 13.3%. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

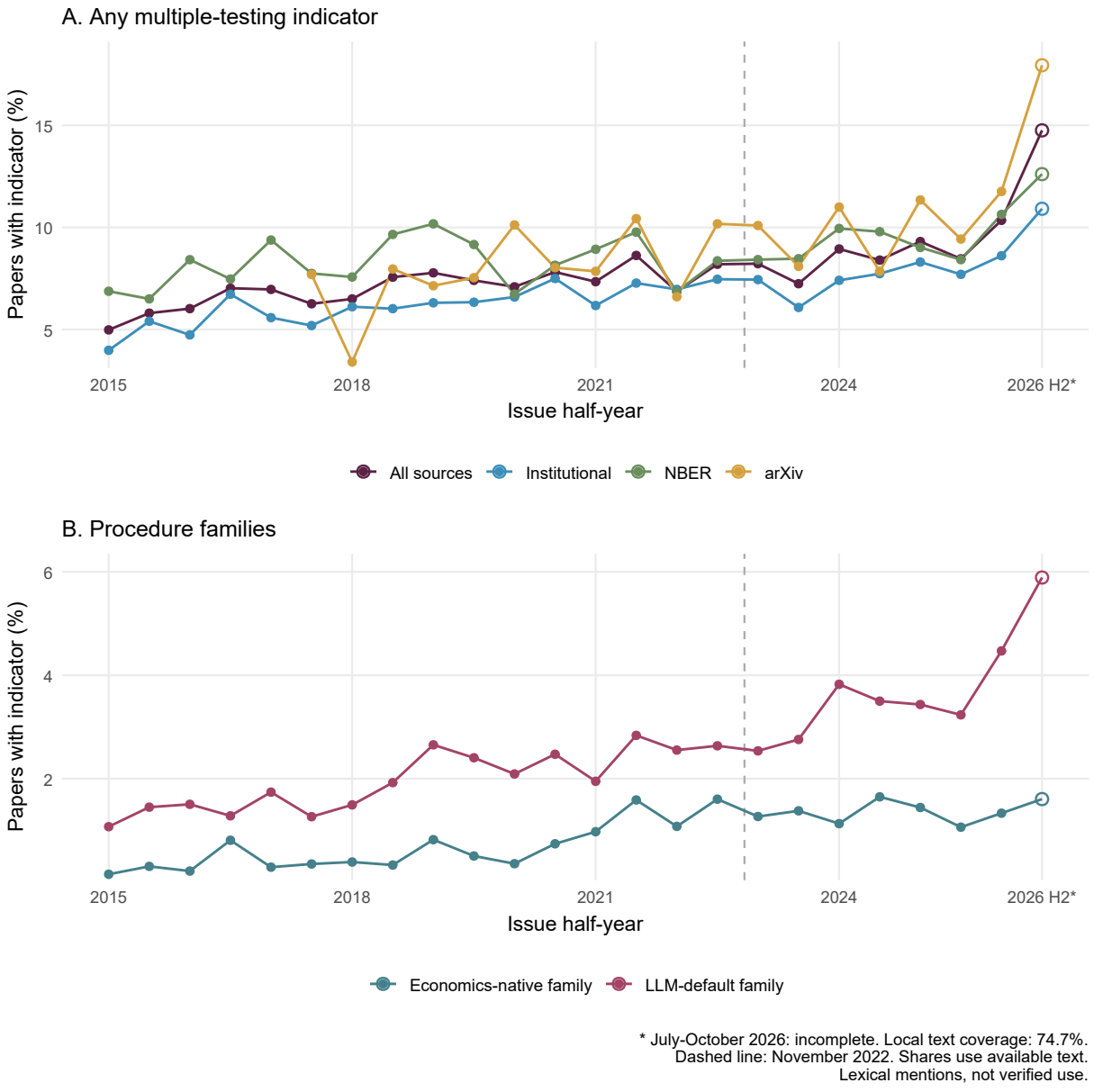


Figure B3. Mentions of multiple-testing corrections

Half-year shares of timing-supported papers with available text whose text mentions a multiple-testing correction outside the reference list. Panel A: any correction by source group. Panel B: the LLM-default family (Holm, Bonferroni, Benjamini–Hochberg) against the economics-native family (Romano–Wolf, List–Shaikh–Xu, Westfall–Young, Anderson). The dashed line marks November 2022. The last cohort is incomplete, with 74.7% text coverage. These are lexical mentions, not verified use. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B20. Correction mentions by half-year: all sources

Half-year	Text papers	Any correction (%)	Holm (%)	LLM-default family (%)	Economics-native family (%)
2015 H1	1,305	4.98	0.00	1.07	0.15
2015 H2	1,310	5.80	0.23	1.45	0.31
2016 H1	1,395	6.02	0.07	1.51	0.22
2016 H2	1,481	7.02	0.20	1.28	0.81
2017 H1	1,379	6.96	0.15	1.74	0.29
2017 H2	1,422	6.26	0.35	1.27	0.35
2018 H1	1,539	6.50	0.06	1.49	0.39
2018 H2	1,507	7.56	0.40	1.92	0.33
2019 H1	1,582	7.77	0.51	2.65	0.82
2019 H2	1,580	7.41	0.32	2.41	0.51
2020 H1	1,960	7.09	0.56	2.09	0.36
2020 H2	2,022	7.81	0.30	2.47	0.74
2021 H1	1,948	7.34	0.21	1.95	0.98
2021 H2	1,762	8.63	0.40	2.84	1.59
2022 H1	1,762	6.87	0.34	2.55	1.08
2022 H2	1,745	8.19	0.17	2.64	1.60
2023 H1	1,812	8.22	0.11	2.54	1.27
2023 H2	1,740	7.24	0.23	2.76	1.38
2024 H1	1,856	8.94	0.75	3.83	1.13
2024 H2	2,000	8.40	0.35	3.50	1.65
2025 H1	2,009	9.31	0.35	3.43	1.44
2025 H2	2,257	8.46	0.31	3.23	1.06
2026 H1	2,774	10.35	0.61	4.47	1.33
2026 H2	1,681	14.75	1.55	5.89	1.61

Timing-supported papers with available text; mentions outside reference lists, not verified use. LLM-default family (Holm, Bonferroni, Benjamini–Hochberg); Economics-native family (Romano–Wolf, List–Shaikh–Xu, Westfall–Young, Anderson). The final half-year is incomplete. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B21. Correction mentions and recorded AI indicators

Group	Papers	Any correction (count)	Any correction (%)	Holm (count)	Holm (%)
All observed	13,225	1,291	9.76	67	0.51
Disclosed	891	123	13.80	12	1.35
Writing flagged	1,103	110	9.97	8	0.73
Operational or mixed disclosure	693	96	13.85	10	1.44
Writing-only disclosure	133	17	12.78	2	1.50
Unclear disclosure	65	10	15.38	0	0.00

Post-2022 papers jointly observed on the indicators; groups overlap. Counts and shares refer to lexical mentions. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B22. Correction mentions: source year estimates (1)

Outcome	Regressor	Coef. (pp)	SE (pp)	Raw p	N
Holm	Operational disclosure	0.742	0.340	.032	13,225
Holm	Writing-only disclosure	0.895	0.706	.209	13,225
Holm	Unclear disclosure	-0.569	0.106	$< .001$	13,225
Holm	Writing flag	0.168	0.251	.505	13,225
Any correction	Operational disclosure	2.160	1.116	.056	13,225
Any correction	Writing-only disclosure	1.459	1.636	.375	13,225
Any correction	Unclear disclosure	3.021	5.557	.588	13,225
Any correction	Writing flag	-0.281	1.126	.804	13,225

The indicators enter jointly; operational and writing-only estimates control separately for unlabelled disclosures. Coefficients and standard errors are percentage points. Prose controls omitted. Source-year and field effects; N counts papers. Standard errors cluster by source-year. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B23. Correction mentions: source year estimates (2)

Outcome	Regressor	Coef. (pp)	SE (pp)	Raw p	N
Holm	Operational disclosure	0.742	0.256	.008	13,225
Holm	Writing-only disclosure	0.895	0.500	.087	13,225
Holm	Unclear disclosure	-0.569	0.083	$< .001$	13,225
Holm	Writing flag	0.168	0.327	.612	13,225
Any correction	Operational disclosure	2.160	1.103	.062	13,225
Any correction	Writing-only disclosure	1.459	1.969	.466	13,225
Any correction	Unclear disclosure	3.021	4.630	.520	13,225
Any correction	Writing flag	-0.281	1.275	.828	13,225

The indicators enter jointly; operational and writing-only estimates control separately for unlabelled disclosures. Coefficients and standard errors are percentage points. Prose controls omitted. Source-year and field effects; N counts papers. Standard errors cluster by source. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B24. Correction mentions: source year estimates (3)

Outcome	Regressor	Coef. (pp)	SE (pp)	Raw p	N
Holm	Operational disclosure	0.722	0.343	.038	13,225
Holm	Writing-only disclosure	0.933	0.698	.185	13,225
Holm	Unclear disclosure	-0.555	0.128	$< .001$	13,225
Holm	Writing flag	0.164	0.254	.519	13,225
Holm	Log words	-0.527	0.680	.440	13,225
Holm	Log windows	0.980	0.605	.109	13,225
Any correction	Operational disclosure	1.883	1.050	.076	13,225
Any correction	Writing-only disclosure	1.979	1.674	.240	13,225
Any correction	Unclear disclosure	3.374	5.454	.538	13,225
Any correction	Writing flag	-0.275	1.106	.804	13,225
Any correction	Log words	-0.550	2.661	.837	13,225
Any correction	Log windows	6.808	2.481	.007	13,225

The indicators enter jointly; operational and writing-only estimates control separately for unlabelled disclosures. Coefficients and standard errors are percentage points. Prose controls included. Source-year and field effects; N counts papers. Standard errors cluster by source-year. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B25. Correction mentions: source year estimates (4)

Outcome	Regressor	Coef. (pp)	SE (pp)	Raw p	N
Holm	Operational disclosure	0.722	0.262	.011	13,225
Holm	Writing-only disclosure	0.933	0.502	.076	13,225
Holm	Unclear disclosure	-0.555	0.096	$< .001$	13,225
Holm	Writing flag	0.164	0.326	.620	13,225
Holm	Log words	-0.527	0.579	.372	13,225
Holm	Log windows	0.980	0.524	.074	13,225
Any correction	Operational disclosure	1.883	0.979	.067	13,225
Any correction	Writing-only disclosure	1.979	1.938	.318	13,225
Any correction	Unclear disclosure	3.374	4.429	.454	13,225
Any correction	Writing flag	-0.275	1.213	.822	13,225
Any correction	Log words	-0.550	3.946	.890	13,225
Any correction	Log windows	6.808	3.723	.080	13,225

The indicators enter jointly; operational and writing-only estimates control separately for unlabelled disclosures. Coefficients and standard errors are percentage points. Prose controls included. Source-year and field effects; N counts papers. Standard errors cluster by source. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B26. Correction mentions: within researcher estimates (5)

Outcome	Regressor	Coef. (pp)	SE (pp)	Raw p	N
Holm	Writing flag	1.098	0.696	.115	15,421
Holm	Disclosed	-0.287	0.638	.653	15,421
Holm	Log words	-0.015	0.920	.987	15,421
Holm	Log windows	0.763	0.813	.348	15,421
Any correction	Writing flag	2.374	1.834	.196	15,421
Any correction	Disclosed	0.650	1.936	.737	15,421
Any correction	Log words	-1.356	3.544	.702	15,421
Any correction	Log windows	6.616	3.400	.052	15,421

Writing and disclosure indicators enter jointly. Coefficients and standard errors are percentage points. Researcher and paper clusters, the original pages-sample expansion and paper-total weights. N counts researcher-paper rows, covering 8,193 distinct papers. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B27. Correction mentions: within researcher estimates (6)

Outcome	Regressor	Coef. (pp)	SE (pp)	Raw p	N
Holm	Writing flag	1.100	0.699	.116	15,421
Holm	Operational disclosure	-0.417	0.616	.498	15,421
Holm	Writing-only disclosure	0.569	2.173	.793	15,421
Holm	Unclear disclosure	-0.810	0.554	.144	15,421
Holm	Log words	-0.020	0.923	.983	15,421
Holm	Log windows	0.769	0.815	.345	15,421
Any correction	Writing flag	2.304	1.793	.199	15,421
Any correction	Operational disclosure	1.349	2.152	.531	15,421
Any correction	Writing-only disclosure	-5.747	4.380	.190	15,421
Any correction	Unclear disclosure	8.677	10.918	.427	15,421
Any correction	Log words	-1.263	3.533	.721	15,421
Any correction	Log windows	6.519	3.387	.054	15,421

The indicators enter jointly; operational and writing-only estimates control separately for unlabelled disclosures. Coefficients and standard errors are percentage points. Researcher and paper clusters, the original pages-sample expansion and paper-total weights. N counts researcher-paper rows, covering 8,193 distinct papers. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B28. Correction mentions relative to the pre-2022 trend

Indicator	Papers	Observed (%)	Trend (%)	Gap (pp)
Any correction	12,577	9.91	9.06	0.85
Holm	12,577	0.62	0.47	0.15
LLM-default family (Holm, Bonferroni, Benjamini–Hochberg)	12,577	4.02	3.33	0.70
Economics-native family (Romano–Wolf, List–Shaikh–Xu, Westfall–Young, Anderson)	12,577	1.36	1.67	-0.31

Linear half-year trends fitted over 2015 H1–2022 H2 and extrapolated to the pooled 2024–2026 sample. Gaps are descriptive deviations, not break tests. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B29. Hit-level precision of correction mentions

Indicator	Reviewed	Valid	Precision (%)	Explicit use
Anderson sharpened q-values	7	7	100.0	7
Benjamini–Hochberg or FDR	20	17	85.0	15
Bonferroni	20	18	90.0	16
Generic phrases	40	24	60.0	19
Holm	50	49	98.0	44
List–Shaikh–Xu	8	8	100.0	6
Romano–Wolf	8	7	87.5	7
Westfall–Young	7	7	100.0	7

Audited snippets, grouped by lexical indicator. Valid correction mentions need not establish actual application; the final column separately counts explicit use. Hit-level precision does not measure recall. These exploratory comparisons sit outside the registered twenty-test family; reported p-values are unadjusted.

Table B30. Step-down inference for main and reference outcomes

Family	Outcome	Indicator	Raw p	RW p	Holm p
Main	Any assistant (pp)	Writing	.155	.819	1.000
Main	Any assistant (pp)	Disclosed	.011	.157	.170
Main	Named assistants	Writing	.236	.898	1.000
Main	Named assistants	Disclosed	.001	.020	.019
Main	Authors	Writing	.863	.998	1.000
Main	Authors	Disclosed	.010	.147	.163
Main	Total pages	Writing	.529	.994	1.000
Main	Total pages	Disclosed	$< .001$	$< .001$	$< .001$
Main	Documented attributes	Writing	.193	.862	1.000
Main	Documented attributes	Disclosed	.055	.488	.772
Main	Solo author	Writing	.299	.921	1.000
Main	Solo author	Disclosed	.829	.998	1.000
Main	Solo, no assistance	Writing	.432	.979	1.000
Main	Solo, no assistance	Disclosed	.917	.998	1.000
Main	Any reader	Writing	.759	.998	1.000
Main	Any reader	Disclosed	.648	.994	1.000
Main	Log named assistants	Writing	.126	.754	1.000
Main	Log named assistants	Disclosed	$< .001$	.004	.004
Main	Capped assistants	Writing	.096	.649	1.000
Main	Capped assistants	Disclosed	$< .001$	.002	.002
Reference	Median reference age	Writing	.044	.140	.131
Reference	Median reference age	Disclosed	$< .001$	$< .001$	$< .001$
Reference	Log reference count	Writing	.338	.487	.565
Reference	Log reference count	Disclosed	.282	.487	.565

Romano–Wolf step-down probabilities use 1,999 null-imposed wild-bootstrap draws, one Rademacher sign per researcher shared across models in each draw, and researcher/paper-clustered studentisation. Holm uses the registered twenty-test main and four-test reference families separately. These registered families are distinct from the exploratory decompositions displayed above, which sit outside the registered family. Neither adjustment covers source shocks.

C Project choice and task savings

C.1 Scope and team size

Equation 1 maximises cooperative surplus over integer $n \geq 1$ and scope $x \geq 0$. Assume $j > 0$, $B(n) > 0$ weakly increasing, $c_o'(a_o) < 0$, and $K(n)$ growing sufficiently quickly to ensure a finite optimum. Credit allocation does not distort design; contributions may differ. The outside option is zero. Comparative statics hold other primitives fixed and condition on production.

At fixed n, strict concavity in scope gives the interior condition

$$\frac{B(n)}{1+x^*} = c_o(a_o) + \frac{jx^*}{n}, \qquad \frac{\partial x^*(n)}{\partial a_o} = \frac{-c_o'(a_o)}{B(n)/(1+x^*)^2 + j/n} > 0. \tag{4}$$

Scope is zero when $B(n) \leq c_o$. For a smooth extension in team size,

$$U_{nx} = \frac{B'(n)}{1+x} + \frac{jx}{n^2} \geq 0, \qquad U_{x,-c_o} = 1, \qquad U_{n,-c_o} = 0.$$

The discrete condition follows directly: for $n' > n$, the derivative in x of $U(n', x) - U(n, x)$ equals

$$\frac{B(n') - B(n)}{1+x} + jx\left(\frac{1}{n} - \frac{1}{n'}\right) \geq 0.$$

Scope and team size have increasing differences, as do scope and cheaper execution. The smallest and largest optimal choices therefore weakly increase as execution becomes cheaper, with discrete rosters changing at thresholds. This result places no sign restriction on comparisons between projects with different benefits, costs, or reporting practices.

C.2 Writing savings and assistance hours

The separable writing cost $C_w(a_w)$ cancels from comparisons between designs, leaving preferred scope and roster unchanged conditional on production. Lower costs can induce entry, changing average observed team size in either direction. Writing costs that depend on scope or the roster can instead change the preferred design. Separability is a maintained case, not a general property of writing assistance.

At an execution price normalised to one, write $e(a_o) = c_o(a_o)$. If all residual execution is delegated, assistant hours are $h = e(a_o)x$. Within a smooth choice region,

$$\frac{dh}{da_o} = e'(a_o)x + e(a_o)\frac{dx}{da_o}. \tag{5}$$

The first term is the saving on existing operations; the second is the work associated with expansion. Hours rise only when proportional scope expansion exceeds the proportional reduction in unit requirements. At roster thresholds the comparison is between optimised levels on either side. Translating hours into headcounts additionally requires hiring, allocation, and crediting rules. The number of acknowledged people does not supply those rules.

C.3 Disclosure cutoff equilibrium

In the model of Section 4.2, a team discloses when b exceeds the incremental reputational cost in equation 3. In a cutoff equilibrium with rational expectations, readers hold correct beliefs about d_n, and the cutoff t_n and disclosure rate satisfy

$$t_n = \frac{s(1-\pi)}{n(1-\pi d_n)}, \qquad d_n = 1 - F(t_n), \qquad nt_n\{1-\pi+\pi F(t_n)\} = s(1-\pi). \tag{6}$$

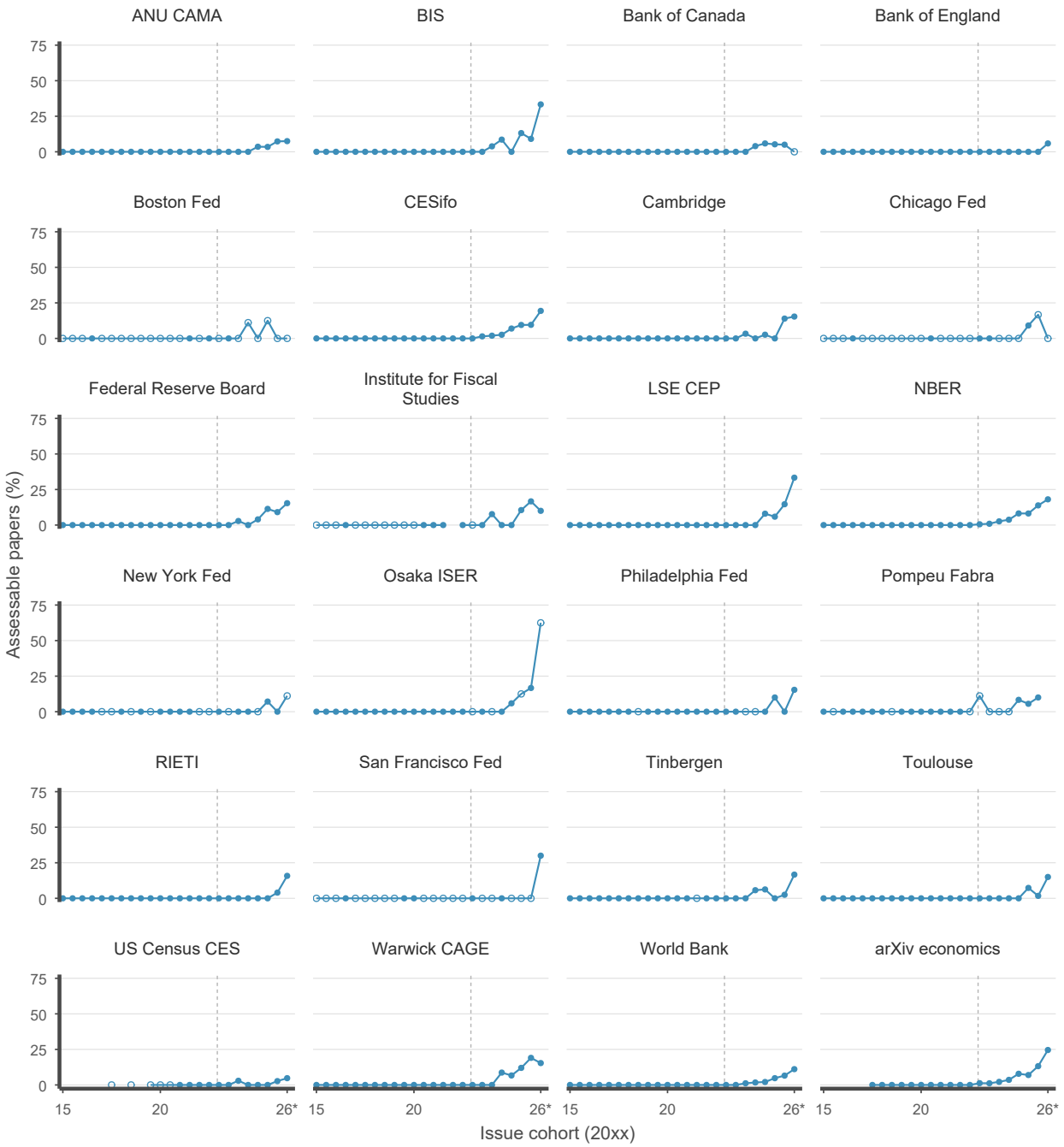
The last expression has a unique positive solution: its left-hand side is continuous and strictly increasing in t_n from zero to infinity. At a fixed cutoff it also strictly increases with n, so the equilibrium cutoff falls as the author team grows. Positive density makes the disclosure rate interior and the comparison strict: $d_{n+1} > d_n$. For a continuous extension in n, implicit differentiation gives

$$\frac{\mathrm{d}d_n}{\mathrm{d}n} = \frac{f(t_n)t_n\{1-\pi+\pi F(t_n)\}}{n\{1-\pi+\pi F(t_n)+\pi t_n f(t_n)\}} > 0. \tag{7}$$

The tie $b = t_n$ has measure zero under a continuous F. As $s \to 0$ the cutoff tends to zero, every user discloses and the size gradient disappears. With k exchangeable operators among n authors, $1/n$ is replaced by k/n in the payoff difference, and the gradient requires k/n to fall with n. These derivations hold other primitives fixed; they do not map onto the fixed-effect regression coefficient, which conditions on researcher, source-quarter, and field effects.

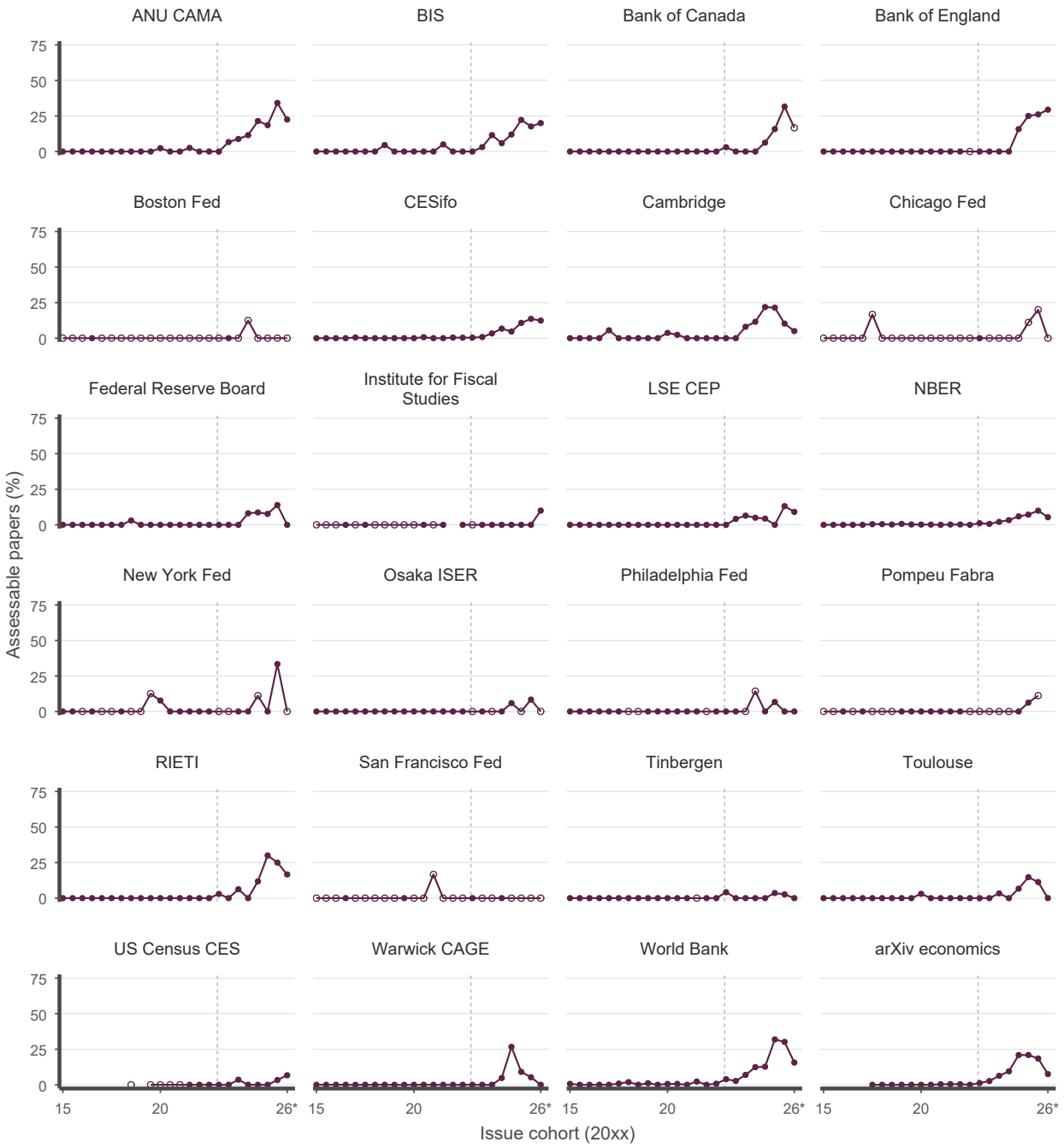
D AI indicators by working-paper series

Figure D1. Disclosed AI use by working-paper series



Half-year issue cohorts, January 2015–October 2026; 26* denotes July–October 2026, an incomplete half-year. Each series uses its timing-supported papers with the indicator assessable, before regression restrictions. Axes are identical across panels and figures. Hollow points contain fewer than ten papers; missing cohorts are gaps. The dashed line marks November 2022. Printed-date observations retain the no-later-revision assumption described in Appendix A; cohort movements can reflect coverage and reporting changes.

Figure D2. Writing flag by working-paper series



Half-year issue cohorts, January 2015–October 2026; 26* denotes July–October 2026, an incomplete half-year. Each series uses its timing-supported papers with the indicator assessable, before regression restrictions. Axes are identical across panels and figures. Hollow points contain fewer than ten papers; missing cohorts are gaps. The dashed line marks November 2022. Printed-date observations retain the no-later-revision assumption described in Appendix A; cohort movements can reflect coverage and reporting changes.